Comparing Interfaces for Image Sorting

A Thesis Submitted to the
College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the degree of Master of Science
in the Department of Computer Science
University of Saskatchewan
Saskatoon

By
Cheralyn Atkins

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan
Canada
S7N 5C9

Head of the Department of Computer Science
176 Thorvaldson Building
110 Science Place
University of Saskatchewan
Saskatoon, Saskatchewan
Canada
S7N 5C9

Abstract

Many tasks ask people to make categorization decisions about a large number of items based on images. There are many ways in which this task can be presented to the user, but there is little information available about how the decision technique affects performance. To investigate this topic three user studies were conducted, which compared different methods for user categorization of images: one based on the "swipe right / swipe left" mechanism commonly utilized in dating sites, four based on grids of images, one that presents a moving stream of images, and one that rapidly flashes a sequence. Additional studies were conducted to investigate the impact of altering certain aspects of the interface design, such as changing the speed of image presentation or adding extra features including zoom and image preview. These studies also investigated the impact of changing the context of the image categorization task, to make the task more familiar to the users. The experimental studies showed strong differences between the various techniques in time, accuracy, effort, and preference, with the small grid the clear winner on all measures. This work provides the first empirical evidence about different approaches for supporting image categorization, and identifies a technique that has distinct advantages over other methods.

Acknowledgements

I would first like to thank my advisors Carl Gutwin and Ian Stavness for all of their guidance and support throughout this process. Secondly, thank you to my thesis committee. Your comments and suggestions have made this thesis better. I would also like to thank the faculty and staff in the Department of Computer Science for all of their assistance throughout the years, as well as all of the members of the HCI Lab for their help and friendship. Finally, I must express my gratitude to my parents and to the rest of my family and friends for providing me with encouragement throughout my years of study and through the process of researching and writing this thesis.

This would not have been possible without all of you.

Thank you.

Contents

List of Tables

List of Figures

List of Abbreviations

ANOVA	Analysis of Variance
CAPTCHA	Completely Automated Public Turing test to tell Computers and Humans Apart
ESP	Extrasensory perception
LOF	List of Figures
LOT	List of Tables
RSVP	Rapid Serial Visual Presentation
TLX	Task Load Index

Chapter 1

Introduction

1.1 Problem and Motivation

There are many situations in which people must make decisions about a set of objects or items based on pictures of those items. Examples of these situations are depicted in Figure 1.1, and can include a variety of specialized or everyday tasks such as choosing t-shirts to purchase from among all of the options available in an online store; choosing images to use in a presentation when performing an image search in a search engine; sorting phenotypic images into groups; "liking" people on a dating site or app such as Tinder [38]; or choosing a subset of plants from a large set of first-generation crosses in a plant breeding project based on their photographs.

These tasks share certain characteristics. They often fit the form of "is this an X?" which could refer to a potential date, good t-shirt design, usable image, successful plant, or desirable phenotype. There are also often many source images to consider (dozens, hundreds, or even thousands). Often it is not possible to automatically categorize these images, and it can be difficult to make bulk decisions from sorting or attribute filtering. In some cases, there are only a small number of True items in the overall set, and there is a predominance of False items (i.e., most things are not an X). Even image recognition tasks that could be automated with a machine learning approach require an initial large training set that must be manually annotated. Finally, people can often pattern-match well based on the image in front of them, and each decision is relatively quick - the task does not usually involve a laborious process of comparing and weighing attributes, and all of the information needed for the decision is usually available within the image itself.

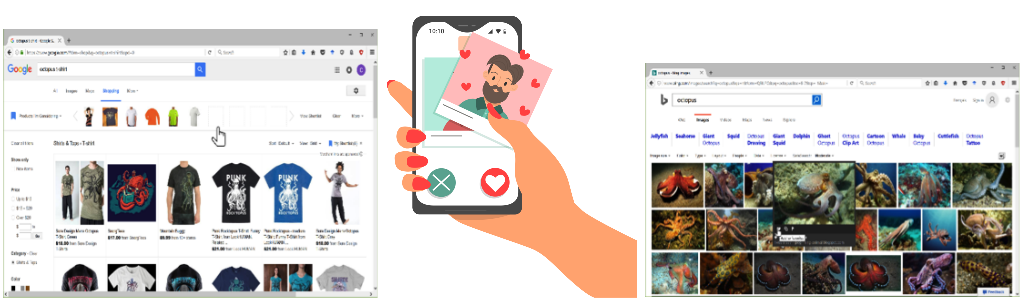

Figure 1.1: Common image-based decision making tasks: Choosing a t-shirt design (left[26]). A swipe-based dating app (middle[21]). An image search engine (right[26]).

Systems designed to support this image-based decision process must obviously use visual presentations of items, but other issues in the task — such as the mechanism for indicating a decision and the details of the presentation — are less clear. It is not known, for example, whether it is better to present a single image at a time, or a set of images; whether people should have a fixed time to decide on each image or an unlimited time; or whether making decisions through tapping, swiping, or checking checkboxes is best (e.g., temporary vs. permanent selection mechanism).

These sorts of tasks becoming much more common — many decisions are being presented through visual means, and many apps are now adopting Tinder-style interaction [31]. It is important to understand how information presentation and selection mechanisms affect the task of image categorization.

1.2 Solution

The objective of this thesis is to discover which method of interacting with images in a dataset was fastest and easiest for image-classification tasks. There were six different research questions that were investigated:

- Which interface style will allow the users to complete the task the fastest?
- Which features of an interface contribute to making it easier or more difficult to use?
- Which type of interaction do users prefer?
- What effect does image speed have in velocity-based interfaces?
- Are certain interfaces better for completing specific tasks?
- Are certain interfaces better suited to specific devices or screen sizes?

In order to answer these questions, three user studies designed to test and compare a variety of interfaces were conducted. Studies were conducted on portable devices, either a tablet or a smart phone, in order to make them practical to deploy outside of a laboratory setting.

The first study had participants complete a fine-grained image categorization task. Participants were asked to categorize images of leaves from the Leafsnap [29] dataset as either matching or not matching the species of the leaf shown in a goal image.

The second study was designed to investigate the impact of changing the amount of time participants had to make categorization decisions. The participants completed the leaf categorization task again, this time with three versions of a selector interface that moved images across the screen in a continuous stream (which will be described in detail in Chapter 3). This study tested three different speeds for the image stream, and compared the participants' performance (speed and accuracy) across the conditions.

The third study involved tasks more in line with what participants may encounter in their daily lives. Participants were asked to look at images showing various scenes and categorize them based on whether or not they contained common objects such as traffic lights. This task is similar to some image CAPTCHAs [40] that people may encounter online. Since this task asked participants to classify images of familiar objects, such as cars and traffic lights, it was expected that the results would show increased user performance. It was also expected that the participants would be able to make the categorization decisions more quickly, as the task did not require specific domain knowledge in the same way that the leaf categorization task did.

1.3 Steps to Solution

There were several steps involved in this investigation. First, the goal was to determine which interface types to test. This was done by investigating common image categorization tasks, and seeing which interfaces are most often used. It was found that many image search engines or online stores display their information in grids of various sizes. This investigation also found that dating apps like Tinder [38] often use a swipe interface to allow users to make decisions. There was also a review of research on different interface designs, such as the Rapid Serial Visual Presentation (RSVP) technique [5] [8] [9], and Interface Currents [24].

Second, key differences between the various interface designs and interaction techniques were identified. These included whether multiple images were presented at once, or shown in series, whether the user or the system determined the pace at which the images were presented, whether designs required explicit actions to identify an image as either a match or a non-match (or assumed all of the images were non-matches by default and the user only had to identify matching images).

Third, the evaluation system based on these key differences was developed, and the first study was performed. This system is described in Chapter 3. Seven interfaces that had participants used to perform a leaf identification task were created. This first study is described in Chapter 4.

Fourth, the impact of speed in velocity-based selection techniques was determined. In the first study, the moving image technique (Sushi selector interface - described in Chapter 3), which presented images in a moving stream, moved quite slowly. This may have caused a few participants to get distracted, and could have affected their task performance. The second study was developed in order to test the impact of changing the speed of the image presentation. This study is described in Chapter 5.

Fifth, the effect of changing the task that the participants were asked to complete was investigated. This was to discover if certain interfaces worked better for completing specific tasks, or if certain interface styles were better suited for specific devices. In a third study, participants were asked to complete a simpler image classification task. Participants were to decide whether or not an image contained a specific everyday item, such as a traffic light, instead of classifying leaves. The device used for this experiment was changed from a tablet, used in the earlier two studies, to a smartphone. This study is described in Chapter 6.

1.4 Evaluation

The main method for evaluating the different types of interfaces was based on the image categorization system that was created. This system is described in detail in Chapter 3, and versions of this system were used in each of the three studies.

The first study had seven interface styles, and each participant completed the leaf categorization task using each of the interfaces. This study used a within-subjects design with all participants. It was expected that differences would be found in the task completion time, and usability ratings between the different interfaces.

The second study tested three different image-stream speeds with the Sushi selector interface. It used a within-subjects design, with all participants completing the image categorization task with all three interface speeds. It was expected that the study would find evidence of a speed-accuracy trade off, with the participants completing the task most accurately while using the slowest interface.

The third study asked participants to complete a CAPTCHA [40] style image classification task with six different interfaces. It used a within-subjects design, and had all of the participants complete the task while using each of the six interfaces. It was expected that the study would find differences in the amount of time required to complete the task with the different interfaces. It was also expected that there would be differences in the usability ratings between the different interface styles.

In all three of the studies the time that it took participants to complete the image categorization task was measured, along with the accuracy of their responses. The participants were also asked to evaluate each of the interfaces using a version of the NASA Task Load Index (TLX) questionnaire [23]. This is designed to measure the workload and subjective amount of effort required for a person to complete a task. At the end of each study participants completed a summary questionnaire, which asked them to rank all of the interfaces that they used in order of their preference.

The results of the first study showed that the Small Grid (4 x 5 images) interface was the most preferred by participants. The four Grid interfaces had the lowest task completion times in this study, and required less than half the time of some of the non-Grid interfaces. The Small Grid also had the best results for accuracy and subjective effort. This interface performed well across all measures, and was the clear winner.

The second study found that the Fast interface was the most difficult to use, and led to less accurate responses. The participants reported feeling rushed, and not having enough time to make decisions. The Slow and Medium interfaces both allowed participants to complete the task more successfully. It was found that the optimal image-stream speed would be between the Slow and Medium interface speeds.

In the third study it was found that the swipe interface performs well with small screens, because the single, large image is easier to see and categorize. It was still found that Grid interfaces perform well, but with smaller screens the size of each image becomes even more important. If there are too many images in the Grid (which leads to each image becoming smaller), the participants may find the images difficult to see and categorize.

Overall, it was found that Grid interfaces are the best for completing image categorization tasks. They allow the users to complete the task quickly and accurately, and require less subjective effort. Designers do need to take the image size and number of images on the screen into account while creating Grid interfaces. It was found that smaller Grids allow for higher accuracy, but may cause slightly high task completion times as well.

1.5 Contribution

The main contribution made by this work is the new understanding of the effects of selection techniques on performance in image categorization tasks. The Small Grid interface is identified as the clear winner in the comparison: this technique's significant advantage in speed and perceived effort is matched by strong performance in accuracy, recall, and precision. These results suggest that designers should strongly consider small grids when developing systems for image-based decision making.

Secondary contributions include:

- Identifying important factors in the task of image categorization (described in Chapter 3).
- Discussing the implications of each of these choices (e.g., providing a larger picture per item vs. being able to compare across several images).
- Providing empirical evidence on how different design factors (number of simultaneous images, time for each decision, need for explicit selections) affect performance, subjective effort, and preference.

1.6 Thesis Outline

Chapter Two provides a literature review of related work in the fields of image classification, human computation, and interface designs. Chapter Three describes the system that was developed to compare possible interfaces for image categorization. The questionnaires that were used in the studies are also described. Chapter Four presents the first of three studies. The purpose of this study was to explore the differences in user experience and task completion time across various interfaces during an image categorization task. Chapter Five presents the second study, which was designed to investigate the impact of different presentation speeds when categorizing images in a moving stream. Chapter Six presents the third and final study, which was designed to investigate whether or not the task that participants are being asked to complete influences their performance or experience with the various interfaces. Chapter Seven discusses some of the overall findings and limitations of the studies and the systems that we developed. It also includes a discussion of possible directions for future work. Chapter Eight concludes this thesis, providing a summary of the findings and contributions.

Chapter 2

Related Work

This research into the effects of various interface designs for completing an image search task was influenced by several previous areas of research. These areas included methods of interactive image selection, automatic classification, human computation, motivation and flow theory, human perception, and psychological myopia.

2.1 Interactive Image Selection

Existing graphical user interface designs that have been used to present large amounts of visual information to users were examined in order to create the interfaces used for this research. Images are commonly presented to users in a grid when performing image searches or in online stores. This is a familiar and fairly straightforward way of presenting information to the user. There have also been some new interface types that have become more popular recently, especially with the increasing prevalence of touch-screen devices and smartphones.

Swipe interfaces have recently become popular with mobile applications [31]. This is partially due to the massive success of the mobile dating app Tinder [38] [31]. This app allows users to view images and profiles of potential partners. If the user wants to select the profile as a match, they swipe right on the image that is displayed onscreen. If the user does not want to match with the profile, they swipe to the left [38].

There has been a lot of research into creating interfaces using the Rapid Serial Visual Presentation, or RSVP, technique [5] [8] [9]. This technique involves rapidly cycling through different images or pieces of text that are displayed on a screen. This can be viewed as the digital equivalent of flipping quickly through the pages of a book. RSVP interfaces trade space for time, and can therefore be useful when using devices with small screens, like tablets or mobile phones [9].

Interface Currents [24] are a type of interaction technique that has been implemented in a tabletop environment. This technique involves graphical representations of files or other pieces of information moving around the tabletop like a conveyor belt. These streams could be repositioned by the users, and information could be added to or removed from the streams. This interface was designed to be used in a collaborative environment, and was useful for allowing users to access hard to reach information [24].

There are a few novel interfaces that have been used to allow users to navigate through image datasets in interesting ways. Cho, Murray-Smith, and Kim [11] compared using tilt dynamics to navigate through an image dataset on a mobile device, to using standard navigation methods such as a scroll wheel or buttons. They found that the tilt interaction method improved the usability of the system [11]. There is also an interface known as "Hands-Free Tinder [37]." This is an app that has been created for the Apple Watch. It is based on the Tinder [38] dating app. In "Hands-Free Tinder [37]," the developers made use of the heart rate monitor in the Apple Watch. Users wear the watch while viewing profiles of potential matches. If the user's heart rate increases while viewing an image, that profile is labeled as a match. If the user's heart rate decreases while looking at an image, it is considered a "left-swipe," or a "no" response [37].

2.2 Automatic Classification

Part of the motivation for this research stems from future plans to create systems that can automatically categorize images. This task can take many different forms, and has several applications. One main application that is of future interest is related to botany. There is work taking place to create systems that can classify images of leaves or plants. These images can be sorted based on different criteria such as plant species, whether or not the plant is diseased, or whether or not the plant has a certain desirable feature or gene.

There are many existing systems that use computer vision techniques to classify images of plants based on specific criteria [1] [4] [7] [29] [32]. To use all these systems, users must first place the leaf against a white background, and take a picture. The image classification process then requires three main steps in order to generate a classification label. The first step is segmenting the image, in order to separate the object of interest from the background. The image is then processed to obtain the relevant features, and assigned a class label using a machine-learning algorithm [1].

Agarwal *et al.* had the goal of creating an electronic field guide that contains all of the plant species represented at the Smithsonian. As a part of this project, they created a system to identify plants based on images of their leaves. The system itself did not solve the problem of leaf identification, but it made it easier for botanists to classify leaves themselves [1]. This system categorized images by looking at the similarity between the shapes of the leaves. The user was shown example images of leaves that cover the 20 best matching species, and was able to choose the closest match from these results. They found that 95% of the time the correct result was returned as one of these 20 matches. The matches were displayed to the users in three different ways, and then they tested the accuracy of the selections. Their results showed that the users were able to select the correct species 70-85% of the time for all of the display methods that were used [1].

Leafsnap [29] is the first mobile app created to classify leaves through automatic image recognition. The app allows anyone to classify leaves easily by taking a picture with a smartphone. The dataset collected for this app contains images of leaves from 184 different species of trees, and is publically available on the

Leafsnap website [29]. Images from this dataset were used in the first two studies that described in this thesis. This app classifies the images using a nearest neighbours algorithm. The extracted features are compared to the database, and the top 25 closest matches are returned to the user. This allows them to make the final choice themselves, by visually comparing the plant in front of them to the images in the app, which cover the leaves, bark, flower, fruit, etc. of each plant. The researchers found that about 97% of image queries returned the correct species as one of the top 5 suggestions [29].

Many of the systems which were designed to "automatically" classify images actually require a human to make the final classification decision. On their own, computer vision techniques are often not accurate enough to properly label image data. Previous researchers have found that incorporating human feedback into an image classification task increased the accuracy of the image labels [7]. As well, a huge amount of labeled data is required in order to train the algorithm in the first place. Therefore, manual labeling is required and is often the only method that allows for obtaining very accurate image labels, however this is expensive and tedious.

2.3 Human Computation

The process of manual image labelling can be improved through crowdsourcing. Crowdsourcing is a method of task completion where a large and often impossible seeming task is broken up into small, manageable pieces, and often up to hundreds of thousands, or even millions, of people online take part [43] [44]. Spreading the task over a much larger group of people means that it will take significantly less time to complete than if a small group of researchers attempted it themselves, and costs can be significantly reduced by having members of the public label the data for free.

Another term often used for projects like these is human computation. This term was first coined by Luis von Ahn, and refers to combining the skills of humans and computers in order to accomplish a task that neither would be able to complete on their own [41]. Computers are able to process vast amounts of information much more quickly than humans, but humans bring some skills to the table that computers are not currently able to replicate. Some of these skills include superior image processing skills, an understanding of the real world, and intuition [36]. These skills have been exploited to try and solve all sorts of problems, from translating the web using Duolingo [41], to trying to identify objects in astronomical images [36].

If someone were to ask for volunteers to help assign labels to every image indexed by Google [26], they probably wouldn't get a lot of responses. One way to convince members of the public to take part in a large, and possibly tedious, task is to present it as a game. This was the goal of the makers of the ESP game, who found that they would be able to accomplish this task if they could get 5,000 users to play the game continuously for 31 days [43]. During the 4 months in which the authors measured usage statistics they generated almost 1.3 million labels for almost 300,000 images [43]. They had almost 400,000 unique players, with 80% of people choosing to play more than once. Some individuals logged more than 50 hours of play time over this time period [43]. The results showed that all of the labels generated by the users made sense in context with the images [43].

Another game, called Peekaboom [44], was created to get users to label object locations in images. This labeling provides data such as which pixels belong to which object, and can be used to create bounding boxes for objects in an image. It can also help computers to locate different objects in an image, or could be used to train machine-learning algorithms [44]. The researchers involved in this project found that the bounding boxes created during gameplay had about a 75% overlap with bounding boxes that were created by hand. There was also the ability for players to "ping" an object, and 100% of the ping locations were determined to be inside of the correct objects [44].

Even users with little to no training are able to accomplish very impressive tasks in a relatively short time. Scientists had been struggling for over a decade to learn about the structure and shape of a specific protein. The challenge was added to the online game FoldIt [12], which asks players to rearrange proteins into the smallest possible shape (their lowest energy form), which is the form that they are most likely to take in nature. Within three weeks the almost 60,000 players had solved the problem, even though most of them had no training in molecular biology [36].

Most people are familiar with *text-based* CAPTCHAs, or Completely Automated Public Turing tests to tell Computers and Humans Apart [40]. These are used by countless websites in order to authenticate users as human. The user is shown a word or a series of letters that has been distorted, and is asked to type the letters that they see. This works because currently, computers are much less capable of deciphering distorted text than humans [42]. These tests are so prevalent, that roughly 200 million CAPTCHAs are typed everyday [41]. Since it takes about 10 seconds for a human to complete a CAPTCHA [41], that amounts to a lot of wasted time.

Figure 2.1: A CAPTCHA example.

The reCAPTCHA [45] project was created to make use of all of that time. These CAPTCHAs are slightly different, and are being used in order to digitize books (usually older books) [41]. The user is shown two words, both of which have been visually distorted. One of the words is taken from the database, so it is a puzzle that has been solved and the computer can tell whether or not the right answer is entered. The other word is taken from an image of a book that has not yet been digitized, so the computer doesn't know what the second word is. The users type both words, without knowing which is which. If the known word is typed correctly, the computer assumes that the user is human, and also assumes that the unknown word is typed correctly. Once a sufficient number of users have agreed on the spelling of an unknown word, that word is considered to have been digitized and is added to the known database [41]. ReCAPTCHA allows users to digitize approximately 100 million words per day, which is equivalent to about 2.5 million books per year. There have been about 750,000,000 unique users, which comes out to about 10% of the world's population [41].

Image-based CAPTCHAs are becoming more common as mobile and touchscreen devices become more prevalent. These types of CAPTCHAs often ask users to look at a group of images and identify or select all of the images that meet a certain goal criteria [10]. These CAPTCHA tasks can also be combined with human computation tasks, in order to leverage the time spent completing security checks. One system shows images of adoptable animals (cats and dogs) from Petfinder.com [35] [17]. Humans are easily able to tell the difference between cats and dogs, while computers have a much harder time correctly classifying these images. The users would be asked to select all of the cat images (for example) from the subset that they were shown. The images would also have an "Adopt Me" link. If the user clicked this link their security check was cancelled, and they were taken to a page where they could adopt the pet shown in the picture [17].

Human computation tasks are the most successful when huge numbers of people take part. These tasks are usually so large that it would be prohibitively expensive to pay people to complete the task, even at very low wages [41]. It becomes important to think about what the users might be getting out of the task, or what might be encouraging them to donate their time and effort, and to keep coming back.

2.4 Motivation

It is important to understand the users' motivation, in order to understand what causes them to devote their time to a specific task. Two major types of motivation are intrinsic and extrinsic. Intrinsic motivation refers to activities that are inherently enjoyable. People take part in an activity solely for the experience, and are not influenced by external rewards. Extrinsic motivation is the term used to describe activities that people take part in in order to earn a reward [16]. When an activity is intrinsically motivating, people will continue to take part in it by choice. They are more likely to return to the activity on their own. Extrinsically motivated actions are often only performed in order to earn the reward. People may even stop enjoying an activity that they previously enjoyed for its own sake if they begin receiving a reward for taking part. When the reward is no longer offered, they are less likely to choose to return to this activity [3].

Flow theory originally stemmed from a desire to understand intrinsically motivated activities, and is often applied to areas of research involving play and games [15] [28]. This theory describes a state in which people become fully immersed in the activity. Flow state can include the following characteristics:

- distortion of time (usually time seems to speed up);
- loss of self-consciousness;
- intense focus and concentration;

- merging of action and awareness;
- the activity becomes intrinsically motivated (the activity itself is a reward) [15].

The theory states that when an optimal balance between the challenge being faced and the player's skill is reached, the user enters a state of Flow. If the activity is too challenging, the user will become anxious and frustrated, and if it is too easy the user will become bored [15]. This relationship is shown in Figure 2.2.

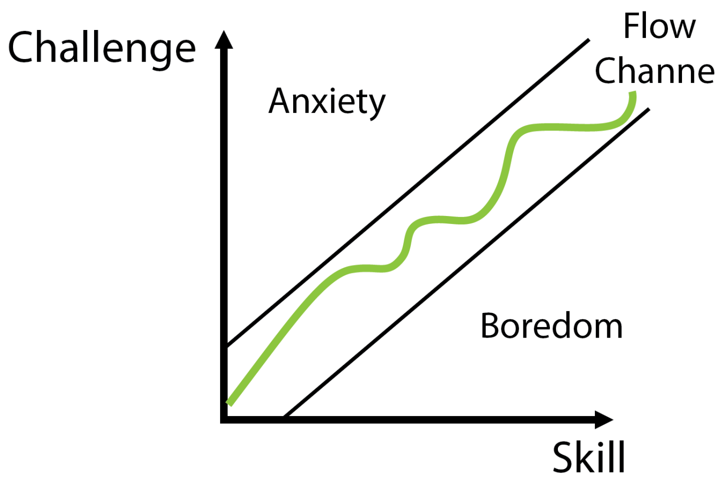

Figure 2.2: The challenge-skill relationship.

2.5 Perception

The distortion of time experienced while in a state of flow can be attributed to problems with human perception. Research has shown that humans are very unreliable at measuring passing time accurately, as they tend to measure it subjectively [19] [27] [30]. There are several factors that can influence or alter how people perceive passing time [2] [18] [20] [30] [34]. There are many examples of illusions or distortions that can cause a person to feel like time has sped up, or slowed down during a certain situation [19].

Subjective time has been modelled as a sort of "counter," which keeps track of a series of internal ticks [19]. One explanation for the apparent distortion of time states that if we imagine that the brain usually processes information at a fairly constant rate, for example, one bit of information is processed per each internal tick of time, then situations in which this processing speed increases could confuse the counter [18] [19]. In times of high stress, such as the moments leading up to and during a traffic accident, a person may be paying more attention to their surroundings and processing information faster, at two bits of information per tick. Since more bits of information were processed in the same amount of objective time, this can trick the brain into thinking that more time has passed. This can be thought to account for the often described illusion that time slows down in moments of high stress [18] [19].

2.6 Psychological Myopia

Psychological myopia refers to the tendency that people have to focus only on information that is immediately related to the choice or judgement that they are making while ignoring background information [25]. One example of this is the money illusion, which shows that in times of inflation people often ignore inflation rate information and base their judgment of a financial outcome on its nominal value rather than on the real, adjusted value. For example, people find a 10% increase in salary when there is a 12% inflation rate more satisfying than a 1% decrease in salary in times of no inflation, even though the latter is better in terms of real monetary value. This illusion shows the tendency that people have to focus on the face value of an event, while ignoring the background exchange rate between the face value and the adjusted outcome [25].

Chapter 3

Design Space and Selection Conditions

3.1 Design Space

The following different design factors were identified that may affect the usability of an interface used for an image categorization task:

- The size of the image,
- Whether single or multiple images are displayed at once,
- Whether or not there is a time limit for each decision,
- Whether or not the user must explicitly categorize every image,
- Whether or not it is possible for the user to correct or change their decision, and
- The physical action required for selection

3.1.1 Size

Different interface types and selection styles require the user to make categorization decisions about images of different sizes. Larger images allow the users to see small details more clearly. This may make it easier for the user to identify the features necessary for the classification process.

The size of the image is influenced by other design factors, such as how many images are displayed at once. Interfaces that only display a single image will have more space available for each image than interfaces with many images. The image size also depends on the screen size of the device being used.

3.1.2 Single vs. Multiple

Interface styles may also differ according to the number of images visible at one time. Some interfaces may present the images to the user in series. This requires the user to make the classification decision based only on the information in the current image. Interfaces that present multiple images to the user simultaneously allow for comparisons between multiple images. This may help the user to identify or rule out images based on similarities or shared features.

3.1.3 Time Limit

Some interfaces allow the user to set their own pace as they move through the dataset. The user can look at images for as long as they want, until they feel comfortable categorizing the image. They can also choose to assign obvious category labels immediately, in order to move through the images more quickly.

Other interface styles may enforce a time limit on the user. At the end of this time period, the image may change or leave the screen. This can force the users to decided on a category label quickly, or else miss their chance. The time pressure may cause users to assign a label before they have enough time to fully examine the image, and could affect the accuracy of their decisions. It could also cause the users to feel frustrated or rushed in their decision making process. The time period may also be too long, forcing users to wait for the image to change after they have already assigned a label. This could cause the users to become bored or distracted.

3.1.4 Explicit Categorization

The user needs to take some sort of action in order to categorize the images. Some interfaces may require both an action to assign a label to matching images, as well as a separate action to assign a label to non-matching images. Other interfaces may assign all of the images a default class label, and the user is only required to select the images that don't belong in this default category. Interfaces that require an explicit action for both positive and negative categories may cause the user to take more time to complete the task.

3.1.5 Corrections

Interfaces also differ in determining when the user's selection is "locked in." In some interface styles, the user may only have one chance to categorize each image. Once the user assigns a class label to an image, that label may be permanent. The user may not be able to go back to previous images and confirm or change their existing classification label. This could decrease the accuracy of the labels if the user assigns a label to the image unintentionally (by "misclicking", for example), or if they erroneously assign a different label than they intended. This may also cause users to become frustrated if they feel that an unintentional action is reducing the accuracy of their selections.

Other interfaces may allow users to correct or change previously assigned labels. The user may be able to "deselect" an image if they change their mind, or had selected it unintentionally. The interface may limit the amount of time after assigning a label that a user has to change it, or may only allow users to reassign labels for a set number of recent images. They may also have a method of moving backwards through the dataset, in order to reevaluate all previously assigned labels. While the ability to make corrections may increase the accuracy of the labels, this backwards movement may slow the task down and increase the amount of time that is required to assign labels to all of the images.

3.1.6 Selection Action

Different physical actions may be required to sort the images into categories. Mobile devices commonly require user to swipe the image in a specific direction [38]. Other touch screen interfaces may require the user to tap an image, or drag it to a specific location. Some actions, especially actions that are already commonly used, may be more familiar to users. Familiar actions may allow users to complete the task faster and more accurately, as they don't have to learn or adjust to new selection techniques. The users may also prefer interfaces that use familiar actions over novel interaction styles.

3.2 Selection Techniques

In order to investigate the differences between interface variations used for image categorization, a system was created that allowed for comparison between techniques for making image-based categorization decisions. Various design factors were incorporated into the creation of this system.

Using these selection techniques, seven different interfaces were designed for use in this study. Four interfaces were based on the Grid technique. These were the Small Grid (4x5), the Medium Grid (5x8), the Large Grid (8x10), and the Zoom Grid (8x10 with zoom feature).The other three interfaces were the Swipe interface, the Sushi selector, and the RSVP interface. These techniques are shown in Figure 3.1, and will be presented in the following sections.

3.2.1 Grid

With the Grid interfaces, several images were presented on the same screen. These images were arranged in evenly spaced rows and columns. Three grid sizes were tested - 20 images (4x5), 40 images (5x8), and 80 images (8x10). A technique where a zoom feature was added to the largest(8x10) size grid was also tested, which allowed participants to double-tap on an image to enlarge it.

All of the images had a grey border when they initially appeared onscreen. This indicated that they were not selected. The participants were able to identify matching images by tapping on them. This would cause the grey border around the image to turn green, which indicated a selected image. The participant could also tap on an image with a green border in order to deselect it if they changed their mind or had made an accidental selection. Once the participant felt that they had selected all of the matching images on a grid page they could tap the next button to move to the next page. There was no way to go back to a previous page. Any images that were not selected by the participant would be categorized as non-matches, while the selected images would be labeled as matching the goal image. The Small Grid and Medium Grid interfaces can be seen in Figure 3.2.

With the Zoom Grid interface, the participant could double tap on an image in order to "zoom in." This would cause that image to move to the center of the screen and increase in size. The goal image and the rest

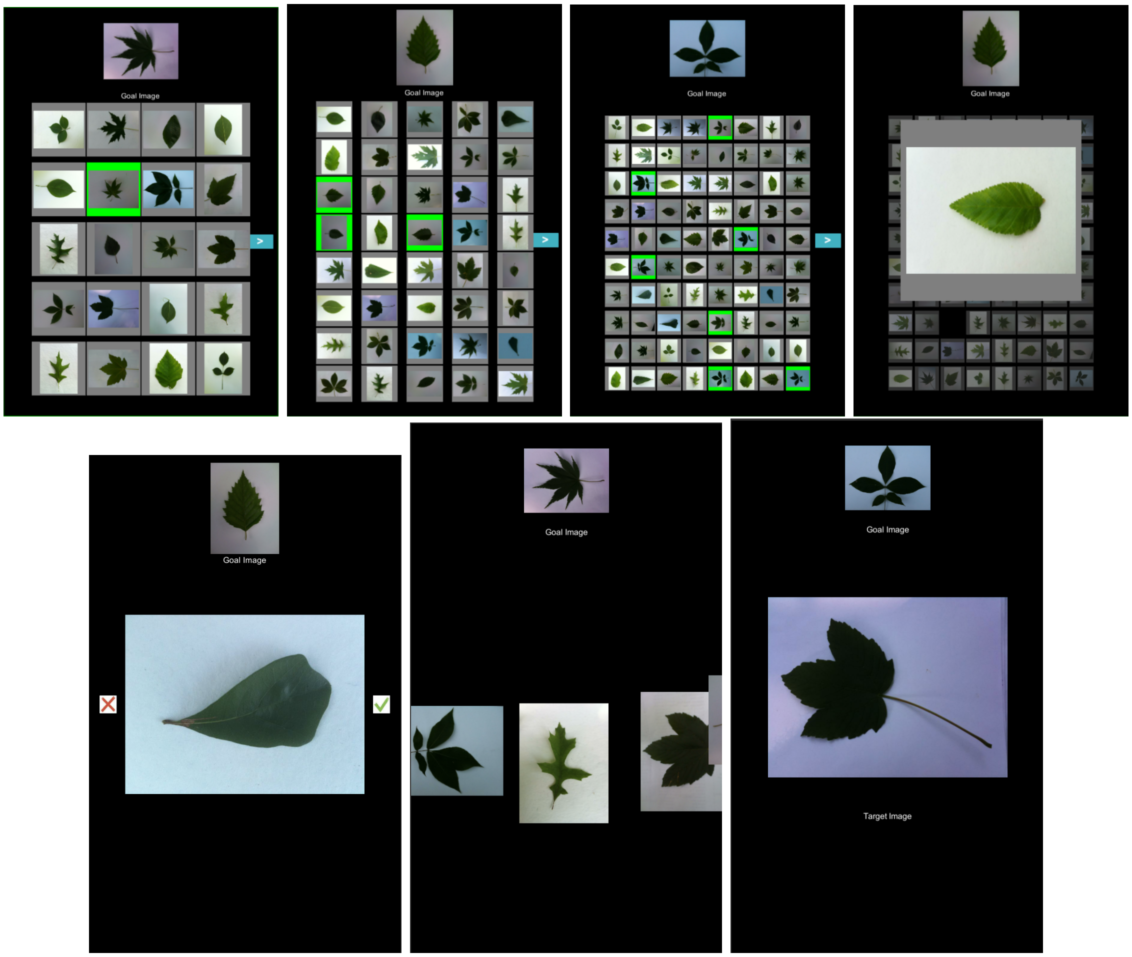

Figure 3.1: Small Grid (far left, top), Medium Grid (left, top), Large Grid (right, top), Zoom Grid (far right, top) Swipe (left, bottom), Sushi (centre, bottom), RSVP (right, bottom)

of the images in the grid would appear slightly greyed out, and would be inactive. The participant could select or deselect the enlarged image by tapping on it. When the participant was finished looking at the image they could “zoom out” by tapping anywhere on the screen outside of the enlarged image. This would cause the image to move back to its previous place in the grid, and return to its normal size. The Large Grid and Zoom Grid interfaces are shown in Figure 3.3.

3.2.2 Swipe

With the Swipe interface, the participant was shown one image at a time. This interface is pictured in Figure 3.4. The participant was presented with a single, large image, and made a categorization image by swiping on the screen. If the participant began dragging or swiping the image to the right, it would tint green to indicate a match categorization. As the participant began moving the image to the left, it would tint red to indicate a non-match categorization.

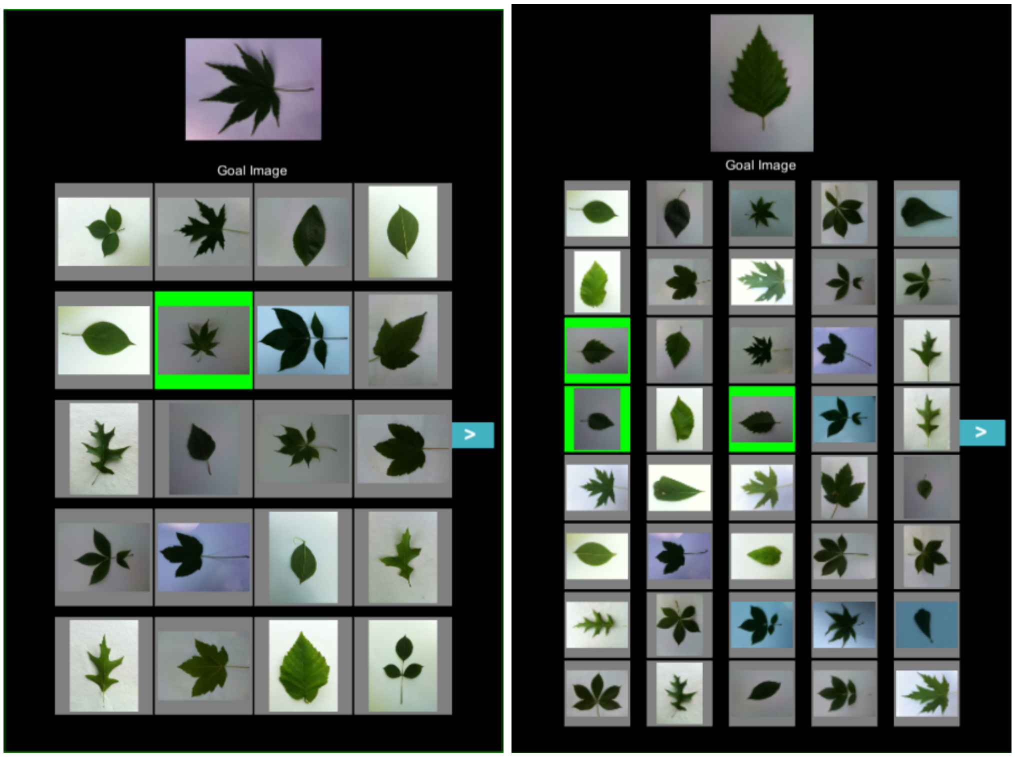

Figure 3.2: Small Grid (left), Medium Grid (right)

An icon of a green checkmark was displayed on the right side of the image and an icon of a red X on the left side of the image in order to remind the participants which image label was associated with each of the directions. When the participant lifted their finger off of the screen at the end of the swipe gesture, the image would be categorized based on the direction that they had chosen, and the next image in the series would be displayed on the screen.

The participant was able to cancel or change a swipe gesture as long as their finger had not left the screen. If the participant began dragging an image to the left, the red tint would appear. If they reversed the drag direction without lifting their finger, they could bring the image back to its original position in the center, causing the colour to return to normal. Removing their finger from the screen at this point would cancel the decision, and the image would not change or be categorized. They could also continue dragging the image past the center, causing the image to tint green. Lifting their finger while the image was tinted green would result in a the image being classified as a match. The participant could not change their decision once their finger had left the screen. There was no way to move backwards through the dataset.

3.2.3 Sushi Selector

The Sushi selector interface displayed a moving stream of images. The images moved across the screen from right to left. There were up to four images visible at one time. This interface is shown in Figure 3.4. Each image was visible for about 5.0 seconds in the first study. In later studies the speed of the image stream was

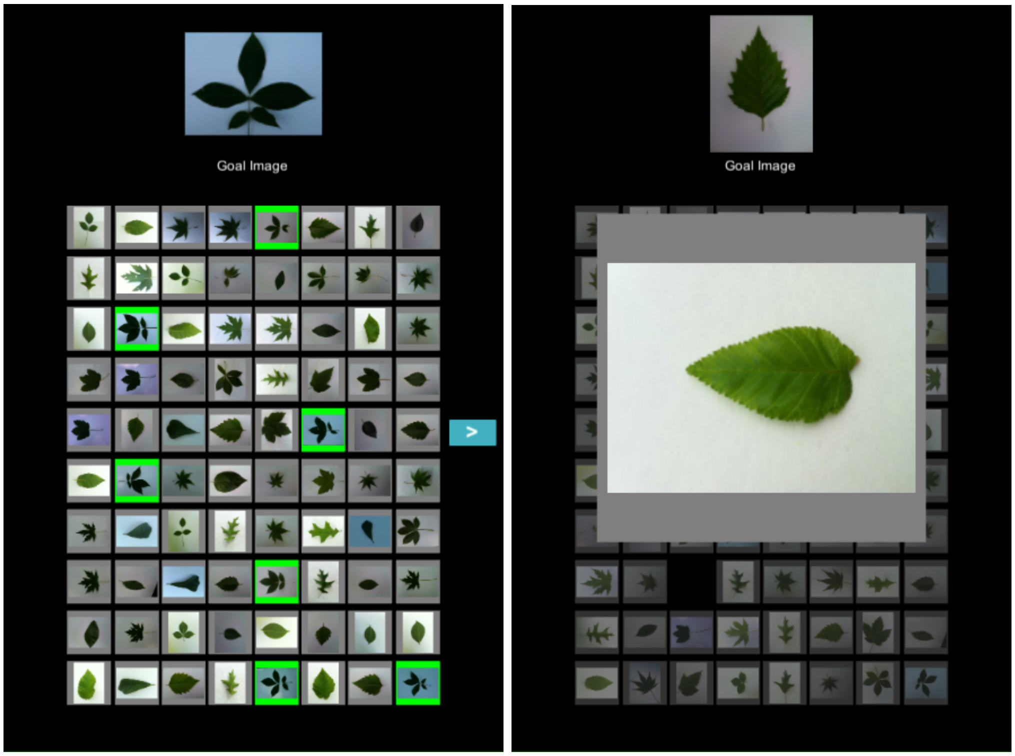

Figure 3.3: Large Grid (left), Zoom Grid (right)

increased.

The participant could touch a matching image and drag it out of line in order to select it. When the participant released the image it would disappear from the screen and be categorized as a match. If the participant did not drag an image, it would continue moving in the stream until it left the screen. Once the image moved off of the screen it would be classified as a non-match. The participant was not able to pause or rewind the image stream.

3.2.4 Rapid Serial Visual Presentation

The Rapid Serial Visual Presentation or RSVP interface displayed a single at a time. Each image was displayed on the screen for one second. After one second the image would change, and the next image would replace it on the screen. In order to select an image, the participant would tap on it. After the participant tapped on an image it would tint green to indicate that it had been selected. If the participant did not tap on an image before it changed, the image would be classified as a non-match. The participant was not able to deselect an image. They were also not able to pause or rewind the image set. This interface is shown in Figure 3.4.

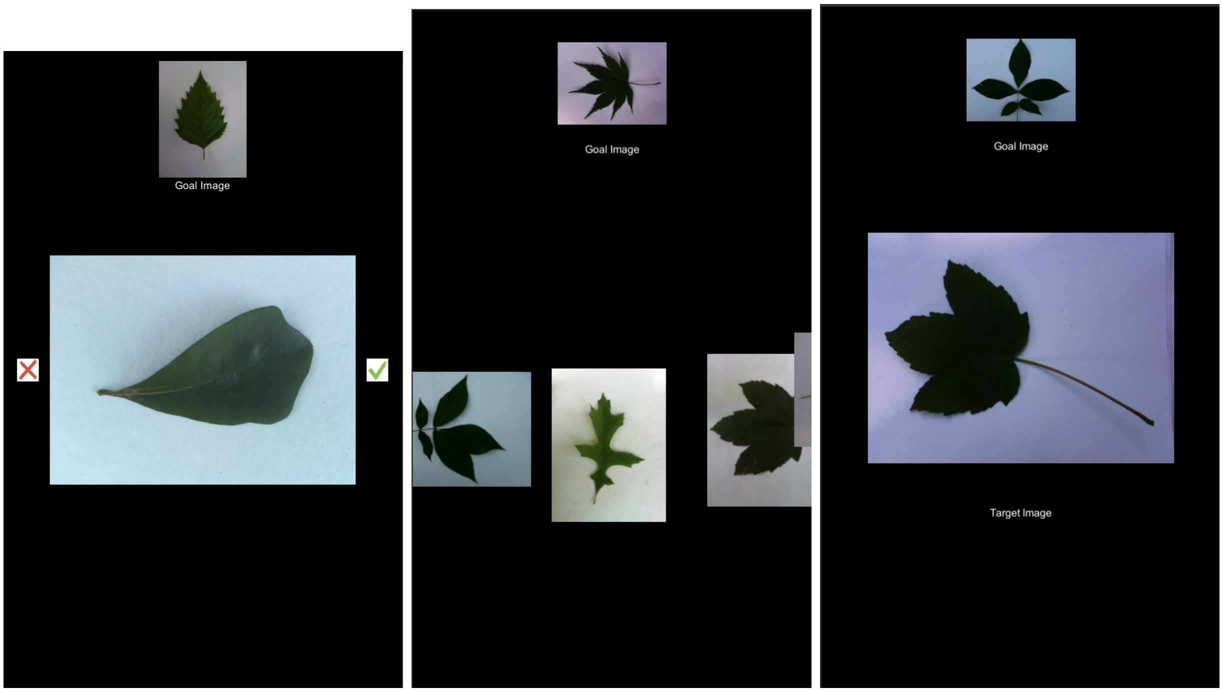

Figure 3.4: Swipe (left), Sushi (centre), RSVP (right)

3.2.5 Summary

These techniques vary on four fundamental dimensions of the design space. The first was whether single or multiple images were presented. The Swipe and RSVP interfaces each showed the images to the participant one at a time, while the other five interfaces displayed multiple images at once. The second was whether a fixed time was provided for each decision. The Sushi and RSVP interfaces both displayed and removed the images automatically with a set time limit. The other five interfaces allowed the participants to move through the image set at their own pace. Third, they varied on whether or not an explicit selection was required for each image. The Swipe interface required the participant to swipe right to select, or left to deselect, an image. They had to take an action to declare each image as a match or a non-match. The other six interfaces required an action to select an image, or declare it as a match, but no action was required to declare a non-matching image. The size of the individual images also varied across the different interface styles. A comparison of the sizes of the images is shown in Table 3.1.

3.3 Devices

This research was intended to develop interfaces that were easy and convenient to use. One of the goals of this research was to investigate which interface styles would best function "in the wild," on devices that people have and can bring with them when conducting daily tasks. For this reason, the focus was on portable devices that used touchscreen interaction.

Table 3.1: Design space of each interaction technique.

Interface	Image Size (on Tablet)	Single vs Multiple	Time Limit	Explicit Categorization	Corrections	Selection Action
Small Grid (4x5)	360 x 360px	Multiple	No	No	Yes	Tap
Swipe	1368 x 1368px	Single	No	Yes	No	Swipe
Sushi	700 x 700px	Multiple	Yes	No	No	Drag
Medium Grid(5x8)	240 x 240px	Multiple	No	No	Yes	Tap
RSVP	1368 x 1368px	Single	Yes	No	No	Tap
Large Grid(8x10)	160 x 160px	Multiple	No	No	Yes	Tap
Zoom Grid(8x10)	160 x 160px	Multiple	No	No	Yes	Tap

The interfaces were originally built to be deployed on a Microsoft Surface Pro 4 with a 12.3-inch display and a screen resolution of 2736 x 1824. This device was used in the first two experiments as a tablet in portrait orientation. It can be seen in Figure 3.5

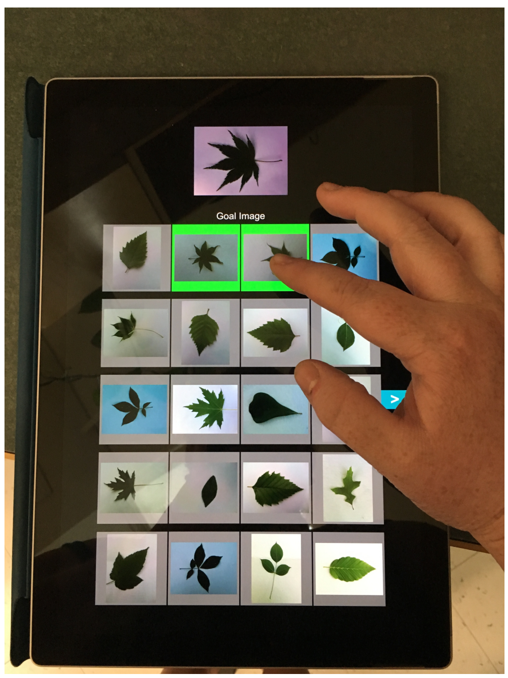

Figure 3.5: System being used on a Microsoft Surface.

A Nexus 5 smartphone running Android version 6.0.1 was used for the third study. This device had a 4.95-inch display and a screen resolution of 1080 x 1920.

All of the interfaces were developed in C#, using the Unity 5 game engine. The software automatically recorded all experimental data. This data included the number of correct responses and errors, as well as selection times.

Chapter 4

Study 1 - Interface Style

The goal of this study was to evaluate styles of interfaces that can be used for image categorization tasks, specifically for sorting images of leaves into groups by species. This study was intended to compare the time it took participants to complete the task with the different interfaces, as well as their performance on the task. It was also intended to examine the user preferences and ease of use of the different interface designs.

4.1 Study Task

In this study, participants were shown a goal image consisting of a picture of a leaf, and asked to select all of the images that contained the same species of leaf. There were three goal images used, each with a corresponding set of 16 matching images. There was also a set of 112 additional distracter images, to add up to a total of 160 images in the dataset. The order in which the participants were shown each of the different goal images was balanced using a Latin square.

The participants were shown the test images in blocks consisting of 80 images per block. Each image was one trial, and the participants were asked to categorize the test images as True or False (matching the goal image, or not matching the goal image). The method of categorizing the images would change depending on which interface style they were using. Each goal image was visible for two blocks, or 160 trials. After completing both blocks, the goal image would change. The same 160 image dataset was used for each goal image, and in all conditions. The task ended when the participant had completed all six blocks, or 480 trials.

4.2 Design

Seven user interfaces were developed for this experiment, which were to be used to complete an image categorization task. The system and interfaces used are described in more detail in Chapter 3.

This study used a within-subjects design, so all of the participants were asked to complete the image categorization task using all seven selection styles. The independent variable was the type of interface used. This study consisted of seven conditions, which referred to the interaction style of the interface. These conditions were: Small Grid (4 x 5); Swipe; Medium Grid (5 x 8); Sushi selector; Large Grid (8 x 10); RSVP; Zoom Grid (8 x 10 with zoom feature). The conditions were balanced using a Latin square to account for possible sequencing effects.

The dependent variables were completion time, accuracy, perceived effort, and user preference. Completion time was measured by the system. A timer counted the milliseconds whenever any of the test images were visible. The timer did not increase between conditions, or if the participant paused between blocks of images. It started counting whenever the participant pressed the button to begin the trials for a block, and stopped when the last test image left the screen.

Accuracy was measured by counting the number of correct and incorrect categorizations. The total number of images categorized using each interface was 480. This was the maximum possible number of correct responses. The total number of correct categorizations that the participant made using each interface was counted, in order to measure their overall accuracy. Precision and recall were also used as indicators of accuracy. Precision refers to the proportion of selected images that were True matches. This is calculated by dividing the number of true positives by the total number of selected images. Recall refers to the proportion of all True images that were selected. This is calculated by dividing the number of true positives by the total number of True images.

Perceived effort was measured using a version of the NASA-TLX [23] questionnaire, which can be seen in Appendix B. This questionnaire breaks effort down into the following components: ease of use, ease of decision making, how overwhelming was the task, how rushed did the participant feel, the participant's subjective accuracy (whether or not they felt like their categorizations were correct), visual appeal, mental demand, and physical demand. The participants were asked to respond to statements that described aspects of these measures using a five point Likert scale from strongly disagree to strongly agree. The participants were asked to respond to the following statements:

- This style was easy to use,
- It was easy to make decisions about a specific image with this style,
- I felt overwhelmed by the number of choices with this style,
- I felt rushed while making decisions with this style,
- I was able to accurately identify matching images using this style,
- This interface was visually appealing,
- Using this selection style was mentally demanding, and
- Using this selection style was physically demanding.

A summary questionnaire (Appendix C.1) was presented to the participants after they had completed the image categorization task with all of the interfaces being tested in that study. The summary questionnaire asked the participants to rank all of the interface styles that they had used in order from best to worst. They were also asked to report which interface they felt allowed them to complete the task the fastest, and with which interface they felt that they had made the fewest errors. This questionnaire was used to get an idea of the overall participant preferences across the different interfaces. It was also used to see which interfaces were preferred overall, and whether this was reflected in the responses from the TLX-style questionnaires or the experiment data.

4.3 Study Conditions

This study consisted of seven conditions, which were balanced between the participants using a Latin square rotation. In all of the conditions, the participants were asked to categorize a set of test images. These images consisted of different types of leaves. Participants were asked to categorize all of the images in which the leaf shown appeared to be from the same species of tree as the Goal image as True matches. Images that showed leaves from a different species of tree were to be classified as False, or non-matches.

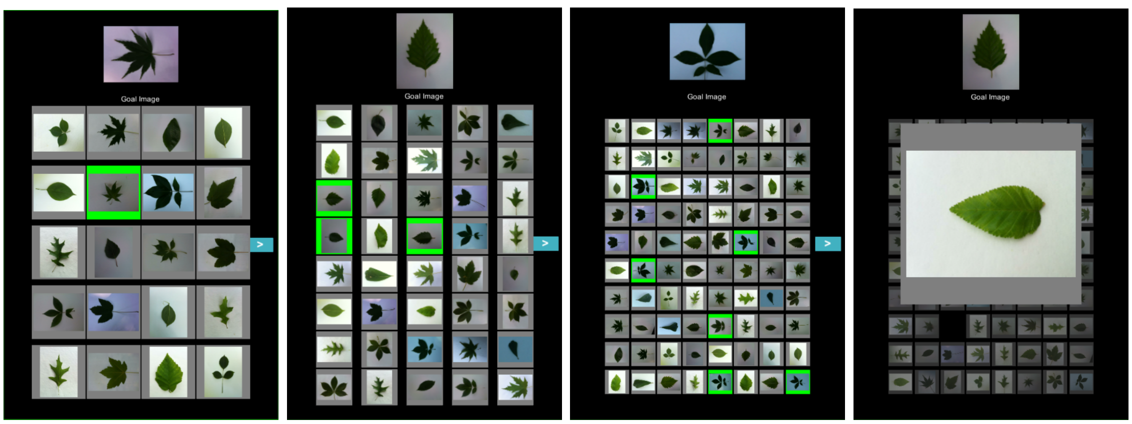

Figure 4.1: Small Grid (4 x 5) (far left), Medium Grid (5 x 8) (left), Large Grid (8 x 10) (right), Zoom Grid (far right - Zoom feature on)

In the **Small Grid (4 x 5)** condition, the participants were shown a Goal image above a 4 x 5 grid of test images, shown in Figure 4.1. They were able to select the images by tapping on them. Selected images were displayed with a green border, so that the participants could keep track of their selections. This can be seen in Figure 4.2. The participant was able to tap again on a selected image in order to deselect it. This gave them the opportunity to correct themselves if they decided that they had selected a non-matching image by mistake.

When the participant thought that they had selected all of the matching images on a page they could push the "Next" button at the right side of the page, in order to advance to the next page of images. If they were on the last page of images in a block, this button would take them to a screen informing them that they were able to pause and take a short break if they wished. When the participant was prepared to continue, they would push the button on this page to resume the task. When completing the final block of images, the "Next" button would change to say "Done" Pushing the "Done" button brought the participant

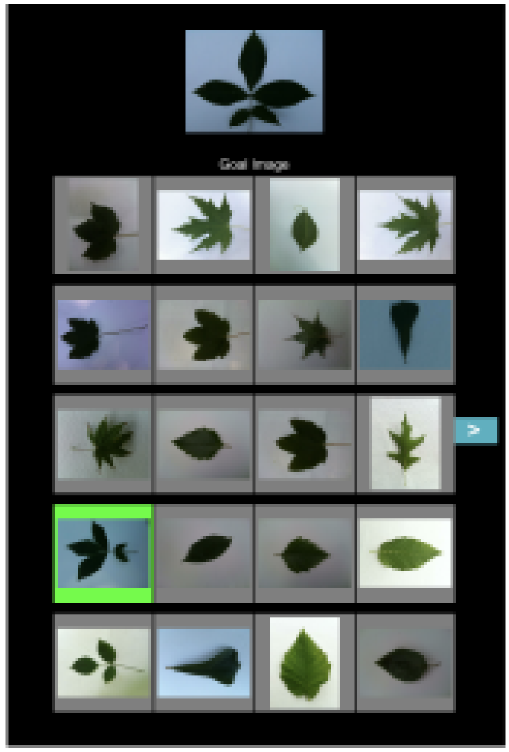

Figure 4.2: Grid showing a green border around a selected image.

to a page asking them to stop and fill out some questionnaires. The buttons can be seen in Figure 4.3, while the break and end screens are shown in Figure 4.4.

The **Medium Grid (5 x 8)** appeared very similar to the Small Grid (4 x 5). In this condition there were 40 images per grid page (5 x 8 images), which can be seen in Figure 4.1. The additional images per page meant that each image had to be smaller in order to fit on the screen. Once again, the participants selected matching images by tapping on them, and proceeded to the next grid page once they thought that they had selected all of the visible matches.

The **Large Grid (8 x 10)** was the third grid-style interface condition. This style consisted of a 8 x 10 grid containing 80 images per page. This interface is shown in Figure 4.1. This meant that each page made up one block of trials. The images in this grid were even smaller, in order to fit on the page.

The **Zoom Grid (8 x 10)** interface was identical to the Large Grid (8 x 10), with the addition of a Zoom feature. This interface also used an 8 x 10 grid of images, that were the same size as the images in the Large Grid (8 x 10) condition. The additional feature allowed the participants to double tap on an image, in order to enlarge it for a better look. This feature is demonstrated Figure 4.1. This was designed to afford participants the opportunity to more closely examine images that may have been difficult to categorize at the smaller scale.

The **Rapid Serial Visual Presentation (RSVP)** interface showed one test image at a time, underneath the Goal image that the participants were required to match. This interface can be seen in Figure

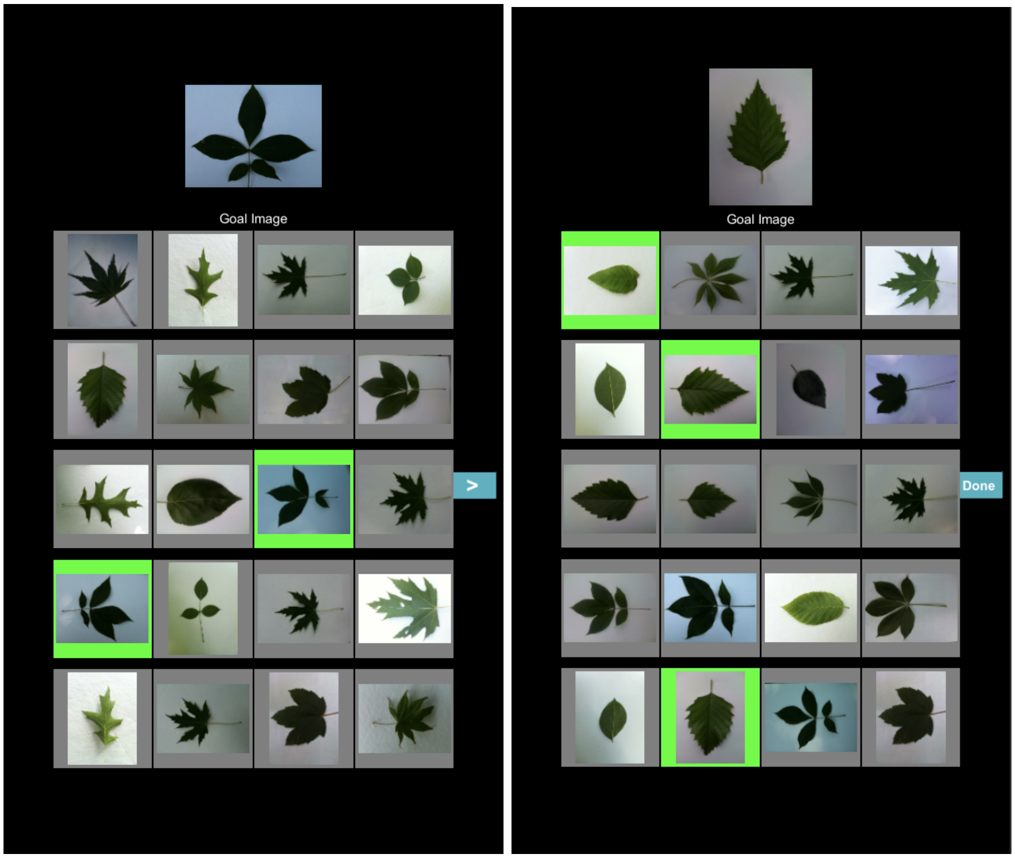

Figure 4.3: Grid interface showing the “Next” button (left) and “Done” button (right).

4.5. Each test image was visible for one second, before automatically advancing to the next image. In order to select a matching image, the participant would tap on the image. Selected images would turn green to indicate the selection, as seen in Figure 4.6. Non-matching images would remain unselected and allowed to advance. After 80 images had been displayed, completing an image block, a button would be displayed and the participant would be prompted to press it in order to continue. At the end of the final block, the prompt would tell the participants to stop and answer questions before continuing.

The **Swipe** interface showed the participants one image at a time, underneath the Goal image. This interface is shown in Figure 4.5. Placing a finger on the test image and dragging it to the left (swiping left) would categorize it as a non-matching image. Swiping the test image to the right would categorize it as a match. In order to remind the participants which direction to swipe for each action, there was a red ‘X’ icon on the left side of the image, and a green ‘check’ icon on the right side. While dragging the image to the right a green overlay would appear to indicate a matching categorization, and a red overlay would appear while dragging to the left. The image overlays are shown in Figure 4.7. After the participant had categorized the final image in a block, a button would appear with a prompt informing them that they could take a break and continue when they were ready. After the final block of images, the prompt informed the participants

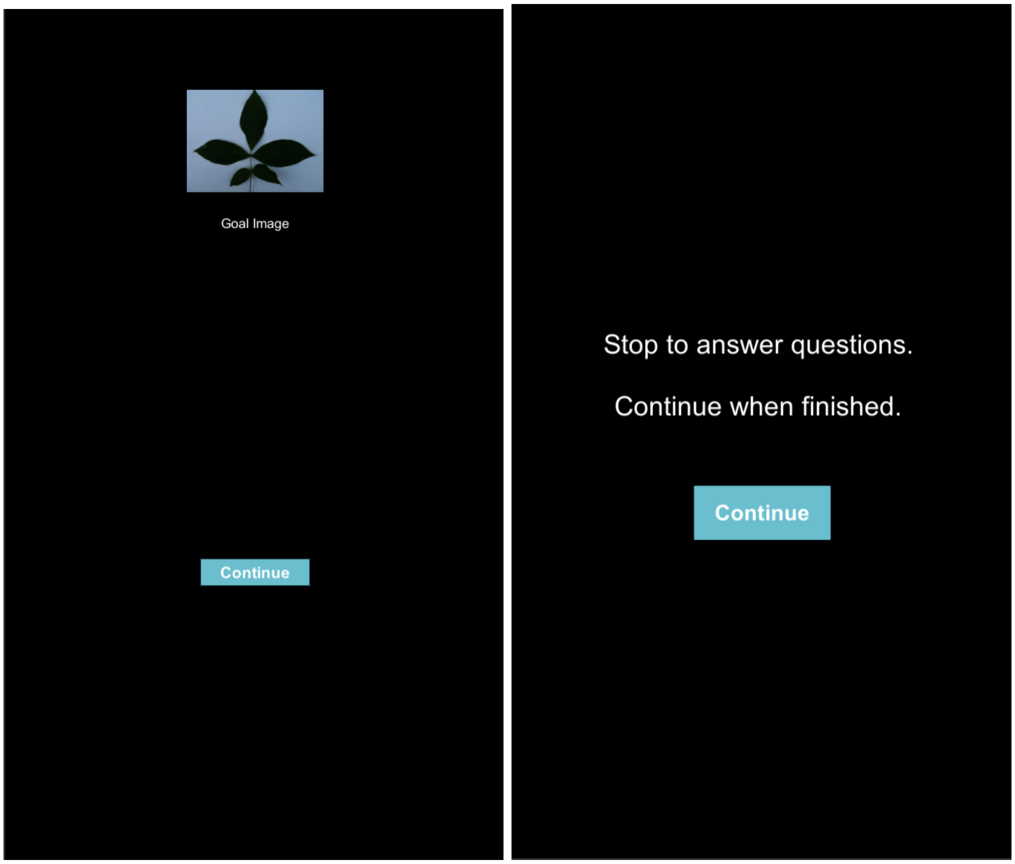

Figure 4.4: Grid interface showing the "Continue" screen (left) and the "End of Condition" screen (right).

that it was time to answer questions.

The **Sushi** selector interface showed the participants a moving stream of test images underneath the Goal image, shown in Figure 4.5. There were up to four images visible at one time, and each image was visible for about 5.0 seconds. The images moved across the screen from right to left automatically. In order to select a matching image, the participant would drag it out of the moving stream. This can be seen in Figure 4.8. Non-matching images would be allowed to advance off of the screen.

All of the Grid interfaces showed the participants multiple test images at the same time, allowing them to compare features across the images. They also allowed the participants to advance at their own pace, and choose when to move on to the next page of images. The participants were required to select all matching images, but didn't need to take any action to categorize non-matching images. Leaving the images unselected caused them to be categorized as False by default. The participants had the option of deselecting images if they felt like they had made a mistake, as long as they had not advanced to a different page of images. The main difference between the Grid interfaces was the size of the images. As more images were added to the page, each individual image needed to be smaller. This made it harder for the participants to see small details in the images. The Zoom Grid interface added the zoom feature, which allowed the participants to

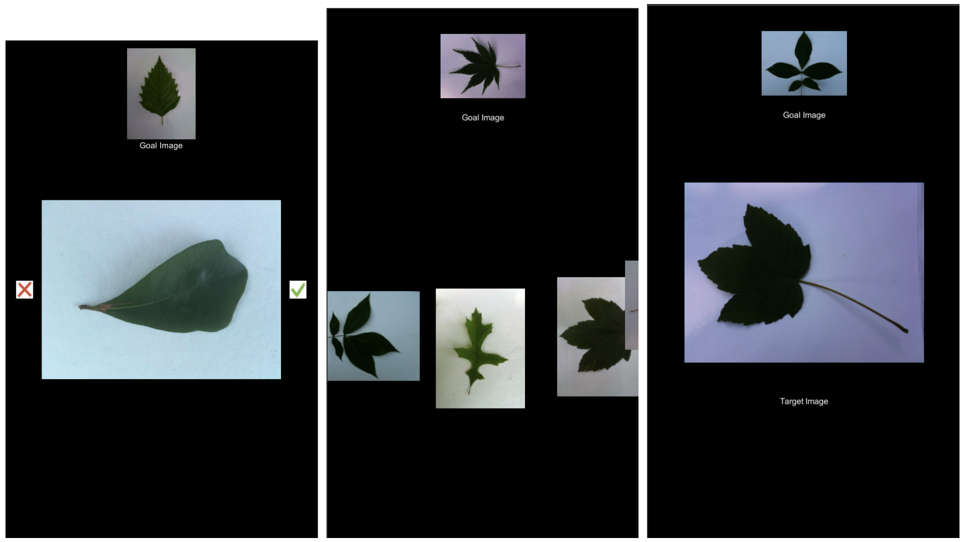

Figure 4.5: Swipe (left), Sushi (centre), RSVP (right)

see a larger version of the image if the wanted to see more details.

The Sushi selector also allowed the participants to compare features across multiple test images. The RSVP and Swipe interfaces only showed one individual test image at a time, so the participants were not able to directly compare multiple images.

The RSVP and Sushi selector interfaces both automatically advanced the images. The participants were not able to control the speed at which the images moved or changed. The Swipe interface moved through the images based on the actions of the participants. They were able to control the speed at which the images advanced, and could move through the dataset at their own pace.

The Swipe interface required the participants to take an explicit action to categorize matching and non-matching images. Swiping to the right categorized an image as a match, while swiping to the left categorized the image as a non-match. The RSVP and Sushi selector interfaces only required an action to categorize matching images. Non-matching images were deselected by default, and did not require further action.

4.4 Procedure

After providing informed consent, the participants were told that they would be asked to complete an image search task using 7 different interface designs. They then filled out a demographics form detailing their sex, any colour vision deficiencies, and how often they use touchscreen devices like smartphones and tablets in their daily lives. This form can be seen in Appendix A.

Next, they began the study with the first of the seven interfaces. They were given instructions on how to use the interface and select matching images, then shown a training block consisting of 80 practice images in order to familiarize themselves with the interface. The practice images were pictures of dogs, and the participant was asked to select all images that contained dogs of the same breed as the goal image (which

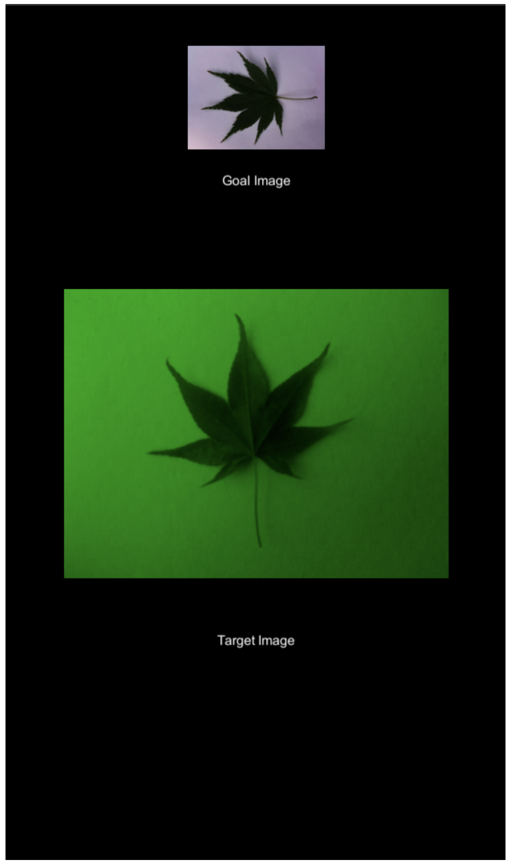

Figure 4.6: RSVP interface showing a "True" categorization.

was a husky). This was followed by 6 testing blocks each consisting of 80 images of leaves. Participants were given a chance to take a break between blocks if they wished. The goal image was changed after the second and fourth blocks, meaning that each goal image was used for 160 trials.

In the Small Grid (4 x 5) interface, there were 20 images per page and the participants flipped through four pages of images for each of the six blocks. The Medium Grid (5 x 8) contained 40 images per page, so the participants flipped through two pages of images per block. In the Sushi selector, Swipe, and RSVP interfaces, the images were presented to the participants in six blocks of 80 images, with the option to take a short break between blocks. The Large Grid (8 x 10) contained six pages of 80 images, with each page representing an entire image block.

The participants were timed while they completed the task, so that it could be determined which selection style was the fastest. The accuracy of the match or non-match labels that the participants provided for each image was measured, so that precision and recall could be calculated for each condition.

After completing all trials, the participants were asked to complete an effort questionnaire based on the NASA-TLX [23] in order to measure the usability of the interface. They then repeated this process for each of the next six interfaces. After completing the final interface, the participants filled out a summary questionnaire asking them to rank the seven interfaces from best to worst, and to report with which interface

Figure 4.7: Swipe interface showing a “True” categorization (left) and a “False” categorization (right).

they thought they performed the task the fastest, and with which interface they thought that they made the fewest errors.

4.5 Participants and Apparatus

This study was performed with 21 participants, all of whom were over the age of 18. There were 14 female participants and 7 male. An advertisement was placed on the University of Saskatchewan student website in order to recruit participants. Most of the participants were students at the university. Seventeen of the participants reported using touchscreen devices, such as smartphones or tablets, for at least one hour per day, with 9 of those participants reporting the use of the devices for more than two hours per day on average.

The study was conducted on a Microsoft Surface Pro 4 with a 12.3-inch display and a screen resolution of 2736 x 1824. The Surface was used as a tablet and kept in a portrait orientation. The software recorded all experimental data, including selection times and errors.

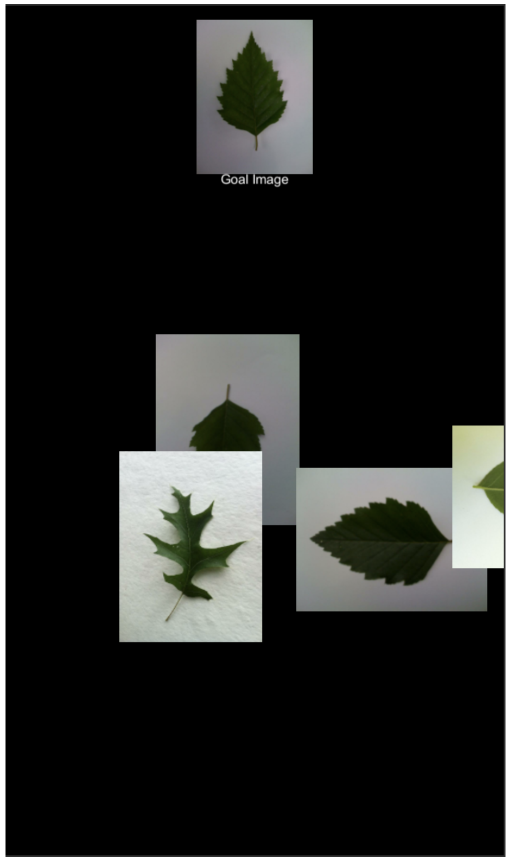

Figure 4.8: Sushi selector showing a "True" categorization.

4.6 Data Analysis

The effort questionnaire that was used provided the following metrics: easy to use, easy to make decisions, overwhelming, rushed, subjective accuracy, visual appeal, mental demand, and physical demand. The participants were asked to rate the items using a 5-point scale from "Strongly Disagree" to "Strongly Agree." The statements "it was easy to make decisions about a specific image with this style" and "I felt overwhelmed with the number of choices in this style" as well as the measures for subjective accuracy and visual appeal were added to the questionnaire in order to check for possible differences between the selection styles. There were also spaces provided for participants to elaborate on their responses or provide additional feedback.

The participants' performance was measured by keeping track of the time that it took them to complete the task using each interface. The system also kept track of the number of images that each participant was able to correctly categorize with each of the interfaces. This allowed for determination of their precision (the proportion of selected images that were True) and recall (the proportion of all True images that were selected) statistics for each of the interface styles. These values also allowed us to calculate the F-measure, which is a balanced combination of the precision and recall values. It is calculated as: F = 2 * [(precision * recall)/(precision + recall)].

RM-ANOVAs were run on the performance and questionnaire data, in order to determine which interfaces showed significant differences. Post-hoc comparisons were performed using the Tukey test. An alpha level of 0.05 was used for all statistical tests.

4.7 Results

For each interface style the time it took for participants to complete the task was measured, as well as the total number of correct categorizations. The number of correct vs incorrect responses was used to calculate the precision and recall values, as well as the F-measure. These measures were used to rate the performance of each participant.

A questionnaire based on the NASA-TLX [23] was used in order to measure the effort required for the participants to complete the task in each of the study conditions. A summary questionnaire was used to determine the participants' interface preferences. RM-ANOVAs were run on the performance and questionnaire data, in order to determine the effects across conditions. The full results of these tests can be seen in Appendix D.

Post-hoc comparisons were performed for all measures using the Tukey test. The full results of these tests can be seen in Appendix E. An alpha level of 0.05 was used for all statistical tests.

4.7.1 Completion Time

Looking at the mean completion time for each interface, shown in Figure 4.9, it can be seen that there are significant results for time differences across the interface styles.

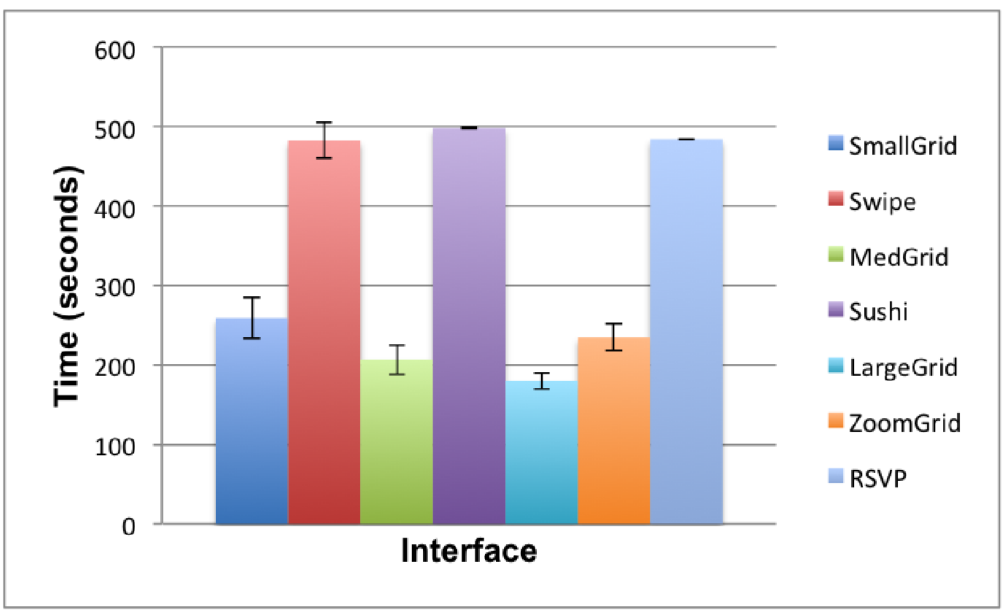

Figure 4.9: Mean task completion time by interface style. Error bars represent the standard error.

It was found that participants completed the task significantly faster with the four grid variations (Small, Medium, Large, and Zoom) than the other three styles. The Large Grid (8 x 10) interface had a mean completion time of 179.33 seconds, with a standard deviation of 46.74 and a standard error of 10.45. The average time for the Medium Grid (5 x 8) condition was 206.23 seconds, with a standard deviation of 81.03 and a standard error of 18.12.The mean time while using the Zoom Grid (8 x 10) interface was 234.16 seconds, with a standard deviation of 75.76 and a standard error of 16.94. With the Small Grid (4 x 5) interface, the participants had an average completion time of 258.58 seconds with a standard deviation of 114.47 and a standard error of 25.60. The Large Grid (8 x 10) had the fastest completion time, and this result was shown to be significantly faster than that of the Small Grid (4 x 5) ($p = 0.015$). There were no other significant differences between the grid variations.

The other three interface styles all had significantly longer completion times than each of the four Grid styles at $p < .05$. The Swipe interface had a mean completion time of 482.16 seconds, with a standard deviation of 102.54 and a standard error of 22.93. The RSVP interface had an average completion time of 483.91 seconds with a standard deviation of .35 and a standard error of .08. The Sushi interface had the highest mean completion time at 497.61 seconds, with a standard deviation of 1.10 and a standard error of .25. There were no significant differences between the Swipe, Sushi and RSVP interfaces.

Table 4.1: Responses to the questions "With which interface were you the fastest?" and "With which interface did you make the fewest errors?" The numbers show how many participants selected each interface as their response.

Interface	With which interface were you the fastest?	With which interface did you make the fewest errors?
Small Grid (4 x 5)	6	6
Swipe	4	3
Medium Grid (5 x 8)	1	0
Sushi	1	3
Large Grid (8 x 10)	3	1
Zoom Grid (8 x 10)	1	4
RSVP	1	0

The participants were asked with which interface they felt that they had completed the task most quickly. These responses are shown in Table 4.1. The Small Grid (4 x 5) was the interface most often selected by participants to be the fastest.

4.7.2 Correct Responses

The mean number of correct responses for each condition is shown in Figure 4.10. These results showed that the mean number of correct responses was highest when using the Zoom Grid (8 x 10), Medium Grid (5 x

8), or Small Grid (4 x 5), but there was no significant effect with an alpha value of 0.05.

There were 480 trials in each condition, making the maximum possible number of correct responses 480. The Zoom Grid (8 x 10) had the highest mean number of correct responses at 453.20, with a standard deviation of 16.88 and a standard error of 4.51. In the Medium Grid (5 x 8) condition the participants averaged 453.05 correct responses, with a standard deviation of 16.85 and a standard error of 4.62. With the Small Grid (4 x 5), the average number of correct responses was 451.90 with a standard deviation of 15.41 and a standard error of 3.08. The mean while using the Sushi selector interface was 450.40 with a standard deviation of 23.03 and a standard error of 5.74. Using the Swipe interface, the participants averaged 446.00 correct responses, with a standard deviation of 20.13 and a standard error of 4.67. The Large Grid (8 x 10) condition had a mean of 446.90, with a standard deviation of 26.55 and a standard error of 4.05. The participants were the least accurate while using the RSVP interface, averaging 436.95 correct responses with a standard deviation of 23.14 and a standard error of 5.55.

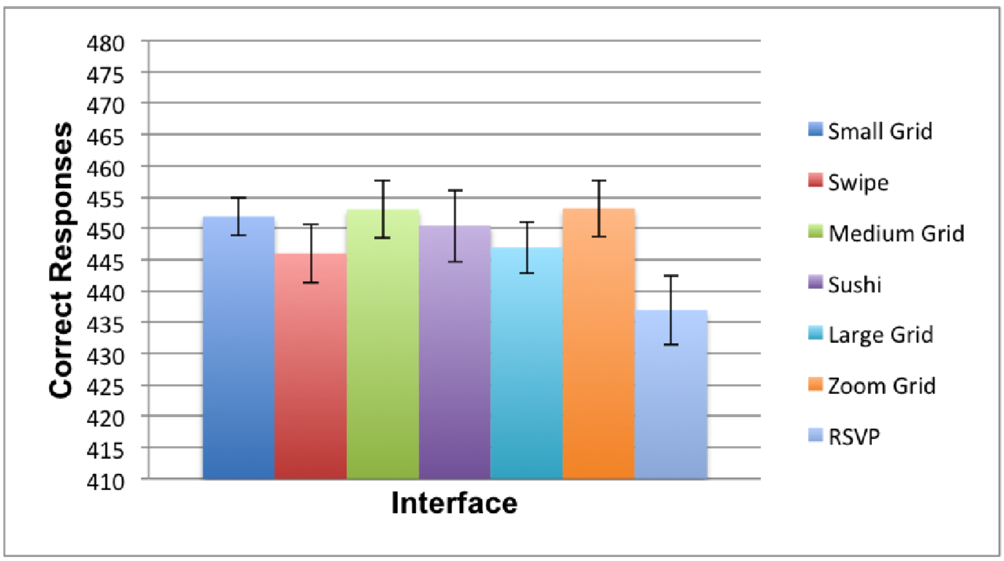

Figure 4.10: Mean number of correct responses by interface style. Error bars represent the standard error. The maximum possible number of correct responses was 480.

Significant differences were seen in precision and recall across conditions. Precision refers to the proportion of selected images that were True matches. This is calculated by dividing the number of true positives by the total number of selected images. The mean precision values for the different interface styles can be seen in Figure 4.11.

The Zoom Grid (8 x 10) condition had the highest mean precision, at .78. The standard deviation was .18 and the standard error was .04. The average precision in the Medium Grid (5 x 8) condition was .75 with a standard deviation of .18 and a standard error of .04. In the Small Grid (4 x 5) condition, the mean precision was .73 with a standard deviation of .19 and a standard error of .04. The mean in the Large Grid (8 x 10) condition was .73, with a standard deviation of .20 and a standard error of .04. The average precision using the Sushi selector interface was .69, with a standard deviation of .19 and a standard error of .05. The Swipe condition had a mean precision of .64 with a standard deviation of .16 and a standard error of .04. Participants were the least precise with the RSVP interface, averaging .58 with a standard deviation of .16 and a standard error of .04.

This research found that precision was significantly lower when using the RSVP interface than it was with the Zoom Grid (8 x 10) and Medium Grid (5 x 8) at $p < .05$. The p-values were .012 (Zoom) and .049 (Med). The Small Grid (4 x 5) interface had the third highest precision score, but there were no significant differences when compared to the other interfaces.

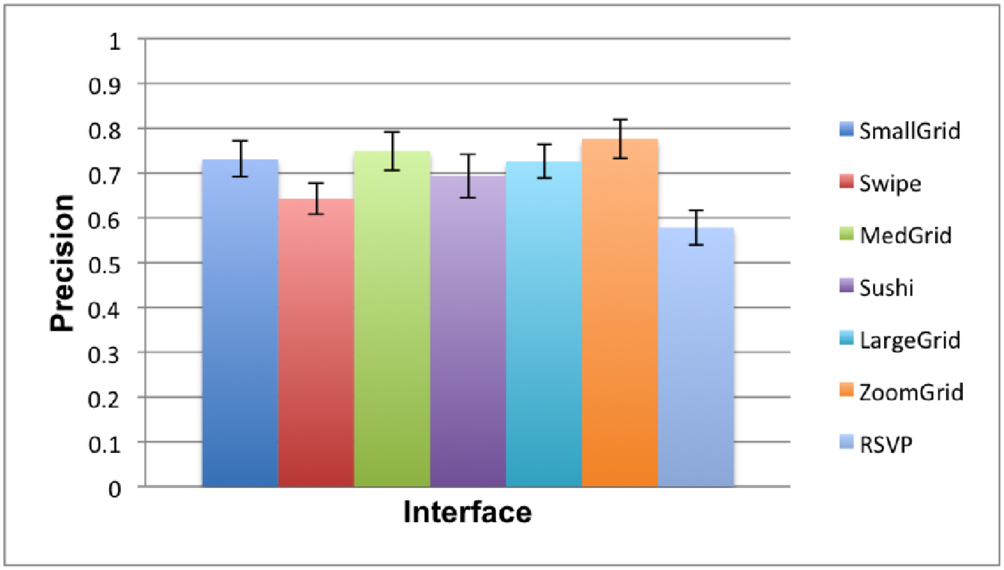

Figure 4.11: Mean precision by interface style. Error bars represent the standard error.

Recall refers to the proportion of all True images that were selected. This is calculated by dividing the number of true positives by the total number of True images. The mean recall values for the different interfaces can be seen in Figure 4.12, as well as in Table **??**.

Using the Sushi selector interface the mean recall was .92, with a standard deviation of .10 and a standard error of .02. In the Swipe condition the mean recall was .88, with a standard deviation of .09 and a standard error of .02. In the RSVP condition the mean recall was .82, with a standard deviation of .15 and a standard error of .04. The mean recall in the Small Grid (4 x 5) condition was .81, with a standard deviation of .15 and a standard error of .04. The average in the Medium Grid (5 x 8) condition was .77 with a standard deviation of .17 and a standard error of .05. In the Zoom Grid (8 x 10) condition the average was .73 with a standard deviation of .19 and a standard error of .05. Recall in the Large Grid (8 x 10) condition was the lowest, averaging .70 with a standard deviation of .18 and a standard error of .04.

Recall was the highest when using the Sushi selector, and these results were significant at $p < .05$ when compared with the Medium Grid (5 x 8), Large Grid (8 x 10), and Zoom Grid (8 x 10) interfaces. The Swipe

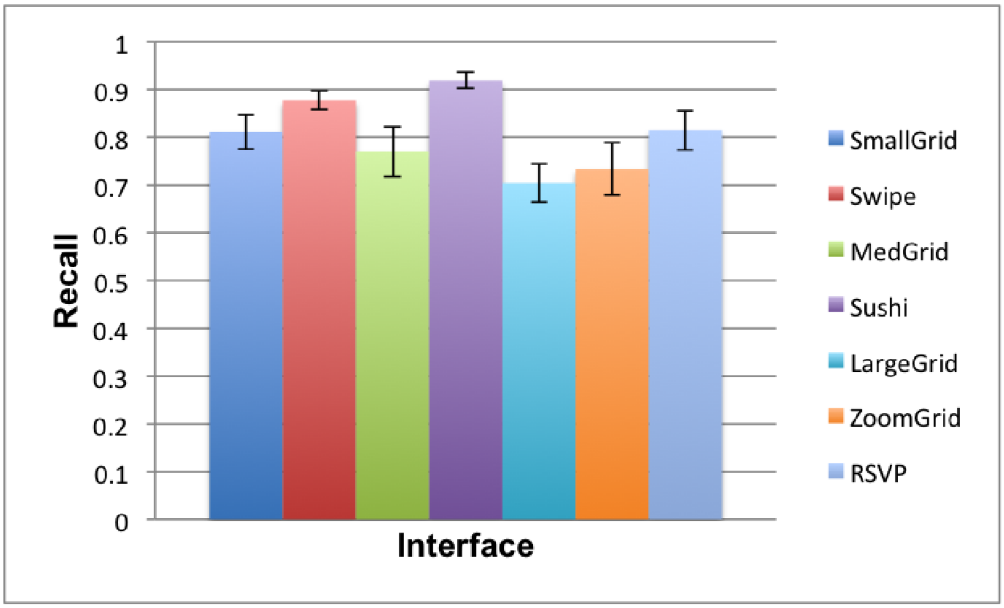

Figure 4.12: Mean recall by interface style. Error bars represent the standard error.

interface also performed significantly better than the Large Grid (8 x 10) interface. The Small Grid (4 x 5) interface had the third highest recall score, but again there were no significant differences when compared to the other interfaces.

The F-Measure is a balanced combination of precision and recall. It is calculated as:

$F = 2 * [(\text{precision} * \text{recall})/(\text{precision} + \text{recall})]$

The mean F-Measure for the different interface styles can be seen in Figure 4.13, as well as in Table **??**. The Large Grid (8 x 10) and RSVP interfaces had a lower F-Measure when compared to the other interface styles. The Sushi and Small Grid (4 x 5) interfaces had the highest F-Measures, although these differences were not significant.

4.7.3 Preference

After completing the image selection task with all seven interfaces, the participants were asked to rank the interfaces from 1 (best) to 7 (worst). Their responses were reverse coded, so that seven would be the highest possible score, and one would be the lowest. This allowed the interface scores to be graphed more easily. The results of this question are shown in Figure 4.14. The Small Grid (4 x 5) was the most highly preferred interface. Swipe, Sushi, and the Zoom Grid (8 x 10) all scored consistently high as well. The Zoom Grid (8 x 10) was preferred over the Large Grid (8 x 10) by the participants. The RSVP interface was the most disliked by the participants.

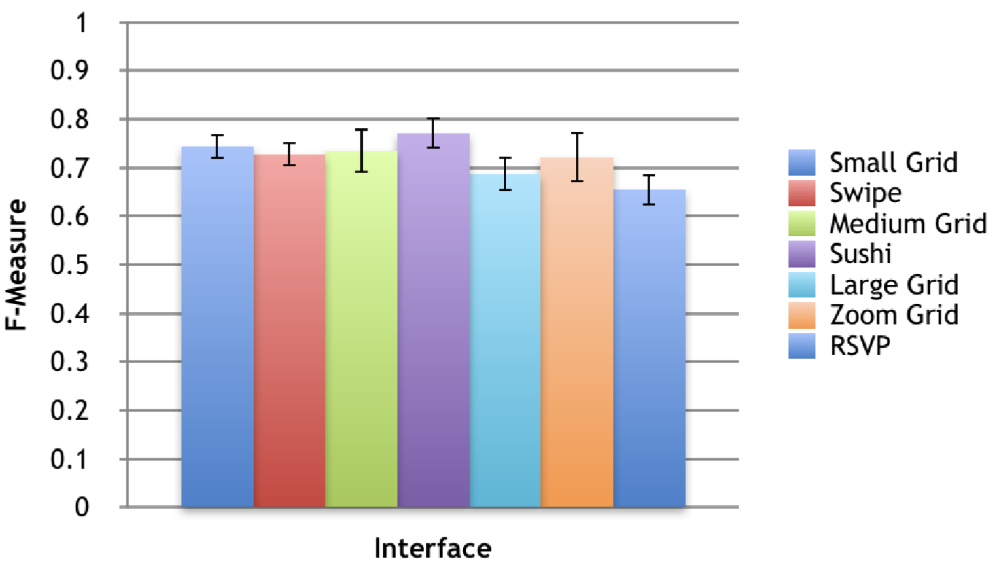

Figure 4.13: Mean F-Measure by interface style. Error bars represent the standard error.

4.7.4 Subjective Effort and Appeal

The levels of mental and physical demand that using the interfaces had on the participants was measured using an effort questionnaire based on the NASA-TLX [23]. The participants were also asked how overwhelmed they felt by the number of choices that they had with each interface, and whether or not they felt rushed. These results can be seen in Figure 4.15. The participants felt significantly more rushed, with an alpha value of $p < .05$, when using the RSVP interface than they did with any other interface except for the Sushi selector (which was not significant with $p = .054$). There were no significant results for physical demand with $p < .05$, although the Large Grid (8 x 10) and RSVP interfaces did tend to be rated slightly higher.

The RSVP interface was rated as being significantly more mentally demanding to use than the Small Grid (4 x 5), Swipe, and Medium Grid (5 x 8) interfaces with $p < .05$. The Large Grid (8 x 10) interface was rated as slightly more mentally demanding than the Small Grid (4 x 5) with $p = .055$, which is not significant. The Small Grid (4 x 5) was found to be the least mentally demanding interface style.

The RSVP interface was found to be significantly more difficult to use than any other interface style except for the Large Grid (8 x 10) with $p < .05$. The Zoom Grid (8 x 10) was found to be significantly easier to use than the Large Grid (8 x 10), and so were the Small Grid (4 x 5) and Swipe interfaces. The participants felt that it was significantly harder to make decisions about a specific image with the RSVP interface than it was with any other interface except for the Large Grid (8 x 10). They also found that it was significantly easier to make decisions with the Small Grid (4 x 5) or the Zoom Grid (8 x 10) than it was with the Large Grid (8 x 10). The Small Grid (4 x 5) interface was reported to be significantly more

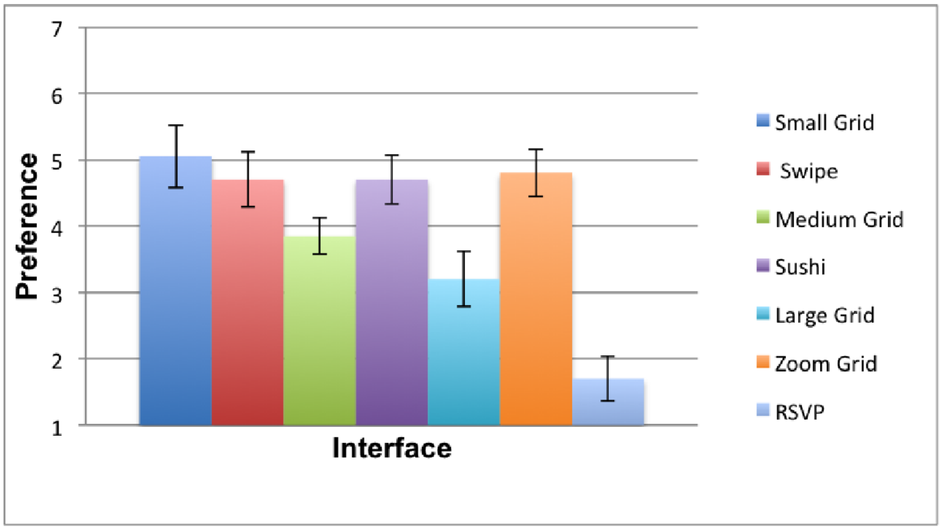

Figure 4.14: Average participant rankings of interfaces (reverse-coded; higher is better). Error bars represent the standard error.

visually appealing than the Large Grid (8 x 10) interface (see Figure 4.15). The participants also felt that their responses were significantly more accurate when using the Small Grid (4 x 5) interface than when using the Large Grid (8 x 10) or RSVP interfaces, with $p < .05$.

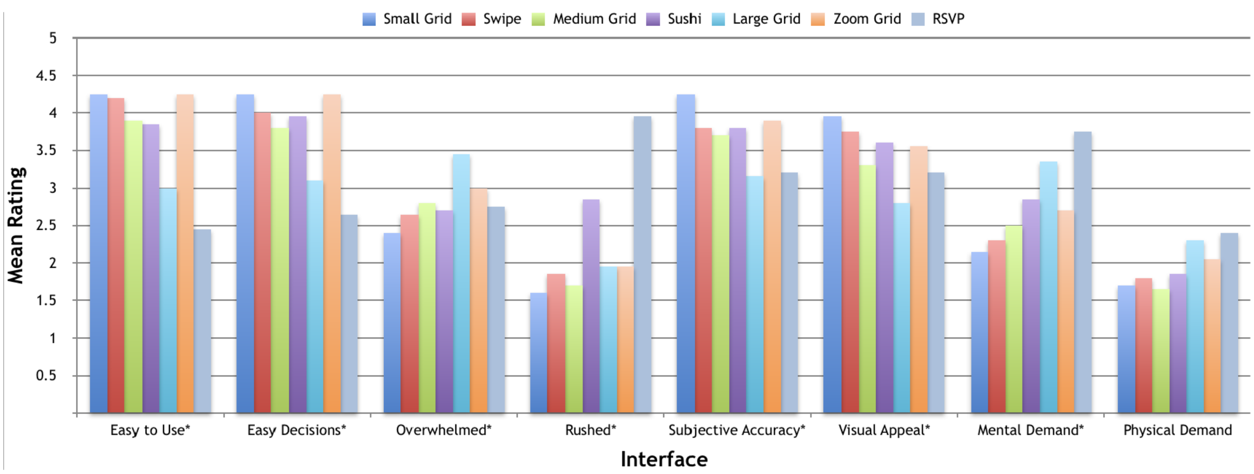

Figure 4.15: TLX-style questionnaire responses by interface style. Measures with significant results are starred.

When asked which interface they felt that they had made the fewest errors with, the Small Grid (4 x 5) was selected by the most participants. These results are shown in Table 4.1.

4.7.5 Participant Comments

At the end of each condition, the participants were asked if they had any comments about the interface or interaction style. In the Small Grid (4 x 5) condition, the participants consistently stated that the interface was easy to use because the images were large and easy to see. They also stated that the categorization decisions were less overwhelming than with the larger grid sizes, since there were fewer images. A few participants noted that the lack of time constraints and stationary images made the task easier to complete at their own pace.

With the Swipe interface, most of the participants found it easier to focus on the single, large image. A few participants found the interaction style confusing, and said that they had a hard time remembering which direction was which. Most of the users found the Swipe action to be simple and intuitive, "just like using a smartphone." Participants said that the "X" and "Check" icons (to indicate "No" or "Yes") made it it fairly easy to figure out which direction to swipe for the desired classification.

The participants found the ability to deselect images very helpful in the Medium Grid (5 x 8) condition. This action was available in all of the grid-based interfaces, but the non-grid interfaces did not allow participants to reverse their selections. Some of the participants commented that there were too many images on the screen at once in this condition, and that the images were too small to see clearly. A couple of the other participants disagreed, and thought that the number of images shown at once was appropriate. The tap-to-select action was thought to be intuitive and easy to grasp.

Some participants thought that the Sushi Selector interface was a bit difficult to use, because it added a time constraint to the task. Most of the comments stated that the images moved slowly, and a lot of time was spent waiting for new images to appear. This wait period could sometimes cause the participants to lose focus on the task. It was suggested that being able to change or select the speed might make the interface more user-friendly. The participants liked that this interface allowed them to compare multiple images, while keeping each image at a large enough size to see details.

The Large Grid (8 x 10) was described as "crowded" and participants found the images to be too small to see details. It was described as straightforward and simple to understand, but the reduced size of the images complicated the task. Some participants described feeling overwhelmed by the number of choices in this condition.

The addition of the Zoom feature (to the 8 x 10 grid) was said to make the task easier. The participants liked being able to zoom in on hard to see images. They were able to confirm their categorization decisions, and felt more confident in their selections. The Zoom Grid (8 x 10) was still described as feeling crowded and displaying too many images at once.

In the RSVP condition, participants felt like the images moved too quickly. They felt rushed, and didn't have enough time to make decisions. The participants also said that sometimes they would attempt to select an image, but the displayed image would change during the selection process and they would accidentally select the next image in the sequence instead.

4.8 Discussion

This investigation into different image selection styles provides five main results:

- The four grid-based interfaces allowed participants to complete the image selection task significantly faster than the non-grid interfaces;
- There were significant effects on precision and recall across interfaces;
- The Small Grid (4 x 5) interface was the most preferred by the participants, and the Swipe, Sushi, and Zoom Grid (8 x 10) interfaces were also highly rated;
- The participants felt more rushed while using the Sushi and RSVP interfaces, and the Large Grid (8 x 10) made them feel overwhelmed;
- There were significant differences in how easy the various interfaces were to use, and how easy it was to make decisions with different interfaces.

The following section provides explanations for these findings.

4.8.1 Explanation and Interpretation of Results

Why were the grid-based interfaces faster than the non-grid interfaces?
There are some key differences between the interaction styles of the different interfaces, which can explain the time difference for the task completion. The first difference is whether or not the participant was forced to adhere to a time limit when reacting to the images. The Sushi and RSVP interfaces both had a time limit, while the other five interfaces did not.

The participants reported feeling more rushed while using the Sushi selector interface than the Swipe or grid-based interfaces, although not to the same extent as with the RSVP interface. With the Sushi selector the images would move from one side of the screen to the other, and move off-screen unless the participant dragged them out of the line-up.

The participants had to make a decision about the image while it was still visible, or else miss their chance. There were some participant comments reporting that they had missed selecting a matching image, especially if multiple matching images were close together. The participant occasionally did not finish the selection action for the first image quickly enough to be able to grab the next matching image before it left the screen. Other participant comments stated that the image belt sometimes moved too slowly, especially if they had already decided that there were no matching images on-screen.

With the RSVP interface, each image appeared on the screen for one second before automatically cycling to the next image. The participants reported feeling significantly more rushed with this style, and had a harder time making decisions about specific images. The participant comments revealed that the image would sometimes change while the participant was making the decision about whether or not to select it. This led to the participant occasionally selecting the next image by accident, or realizing that they had missed a matching image.

What is interesting about this is that the participants completed the task more quickly with all four grid-based interfaces than with either of the timed interfaces. On average, it took the participants about twice as long (or more) to complete the task with the non-grid interfaces. A possible explanation for this comes down to another difference between the interface styles. This was whether the participant was able to view only one image at a time, or was shown multiple simultaneous images.

With the four Grid interfaces, and the Sushi selector, there were multiple images on-screen at a time. With the Swipe and RSVP interfaces there was only one image visible at any point. When there were multiple images showing simultaneously, the participants were able to compare the images to each other as well as to the goal image. The participant comments showed that this helped them make decisions about some of the images.The participants were also able to scan over or process multiple images simultaneously. This meant that the participant could be making a decision about another image (or multiple other images) while they were in the process of taking an action to select a specific image. This decreased the average time required to make a decision about each specific image individually. Because the decision and selection process could take place simultaneously for multiple images, the participant was able to complete the task more quickly overall.

The Swipe interface differed from the rest because it required the participant to react to every single image. In the other six interface styles (Grids, Sushi, RSVP) the participant was required to take an action in order to categorize an image as True, but taking no action would cause an image to be labeled as False by the system. With the Swipe interface, the participant had to swipe right in order to label the True images, but they also had to swipe left in order to label the False images. Taking no action would stop the progression of images. The participants reported that the Swipe interface was easy to use and allowed them to make decisions easily. In spite of this, it took almost twice as long to complete the task with the Swipe interface than with any of the four Grid interfaces. This is because the slight amount of extra time that was required to take an action for each False image added up to a significant amount of time overall.

Did the selection style affect the number of correct responses?

This study did not find any significant differences in the total number of correct responses across the different interface styles. The data trends showed fewer correct responses with the RSVP, Swipe, and Large Grid (8 x 10) interfaces.

There were significant result in regards to precision and recall. Recall was the highest with the Sushi selector interface. The Swipe, Small Grid (4 x 5), and RSVP interfaces also performed well for this test. The Large Grid (8 x 10) and Zoom Grid (8 x 10) interfaces had the worst performance for recall. Precision, on the other hand, was the highest with the Zoom Grid (8 x 10) interface. All four of the grid-based interfaces performed well for precision. The least precise interfaces were the RSVP and Swipe interfaces.

When performing some types of image categorization tasks, there is often more emphasis on Precision than there is on Recall. This can be because the large number of source images often means that missing some of the True images isn't as much of a problem as including some False images. With tasks such as online shopping, using a dating app, or choosing images from a Google-Images [26], including false positives would be much more of a problem for the user than missing some of the matching images. A false positive in these situations could mean accidentally buying an unwanted item or using an image that doesn't make sense in a presentation, while a false negative would simply mean not buying an item that they may have liked or using a different relavent image in the presentation. This isn't necessarily true when sorting objects into categories, because the False category may be as important as the True category.

How did the size and number of images on the screen affect the use of the interface?
Comparing the different grid styles exposes some of the clearest results that altering the size and number of images has on the use of the interface. This research found that the Large Grid (8 x 10) was significantly harder to use than the Small Grid (4 x 5), and made the task more mentally demanding. The participants found it more difficult to make decisions with the Large Grid (8 x 10), and reported feeling more overwhelmed when using the Large Grid (8 x 10). The participants preferred the Small Grid (4 x 5) interface over the other grid styles, and felt that the Small Grid (4 x 5) was more visually appealing. The participants also reported that they felt like they were more accurate when using the Small Grid (4 x 5) interface. In all of these measure the Small Grid (4 x 5) performed better than the Large Grid (8 x 10), and the Medium Grid (5 x 8) was found to be somewhere between the two. This indicates that adding more images per page (causing the images to decrease in size in order to fit) can be a problem for the participants. The participants completed the task more quickly with the Large Grid (8 x 10), but they found it more difficult. Even though the number of correct responses wasn't significantly affected, the participants felt that the Large Grid (8 x 10) interface required more effort, and they preferred to use the Small Grid (4 x 5).

Did the Zoom feature improve the use of the Large Grid (8 x 10) interface
The Large Grid (8 x 10) and Zoom Grid (8 x 10) interfaces had the same number of images per page, and the same layout, but with the Zoom Grid (8 x 10) the participants could double-tap an image in order to see a blown-up version of that picture. This study found that the participants preferred the Zoom Grid (8 x 10) interface over the Large Grid (8 x 10) interface. They also reported that the Zoom feature made the interface significantly easier to use, and allowed for easier decisions about specific images. Our results showed some slight trends towards the Zoom feature improving the overall number of correct responses as well as the precision and recall, but these differences were not significant. It was also found that the Zoom interface was significantly slower than the Large Grid (8 x 10) interface. This shows once again that if the images on the screen are too small it can make the participant feel like the interface requires more effort to use, and can lead to them disliking the interface.

Which interface was the best overall?
Most of the interfaces had their pros and cons. For example, the Large Grid (8 x 10) interface allowed

participants to complete the task quickly, but it could be overwhelming and hard to use. The one interface that performed well across all of the categories was the Small Grid (4 x 5) interface. This interface allowed participants to complete the task fairly quickly, although it wasn't the absolute fastest. The participants also found that it was easy to use and visually appealing. There were no significant differences in the number of correct responses when comparing the Small Grid (4 x 5) to the other selection styles. The Small Grid (4 x 5) was also the interface that was most often preferred by the participants. For these reasons, the Small Grid (4 x 5) interface is the clear winner in this experiment.

4.9 Summary

Participants are often required to filter or categorize a large number of images based on certain criteria. This can be true in several different contexts including shopping, or online dating. There is little existing information about how the selection technique affects the participant's performance on the task.

In this chapter, a study was described in which participants were required to complete an image categorization task using seven different image selection styles. A controlled study with 21 participants showed that there were significant differences in the time required to complete the task, number of correct of responses, subjective effort, and participant preference across the different interfaces. It was found that the Small Grid (4 x 5) interface performed well across all of these categories, and was the most preferred by the participants. This interface was the clear winner based on the criteria of this study.

The research into the different interface styles gave strong initial results, and interesting avenues for future work. This study led to interest in exploring the impact that the interface style has on different tasks. It was determined that the next step was to investigate whether certain interface styles are more suited to performing specific tasks, and whether this affects the participant's experience or task performance. The results of this research are presented in the following chapters.

Chapter 5

Study 2 - Image Stream Speed

Following Study 1, there was a research question related to the speed at which the images were shown to the participants. The results from Study 1 showed that the four Grid interfaces, where the participants were able to set their own pace and view multiple images at once, were significantly faster for completing the task. The RSVP, Swipe and Sushi interfaces showed no significant differences in completion time between them. The Swipe interface allowed participants to set their own pace, while only viewing one image at at time. The completion time for this interface varied depending on the participant. The RSVP and Sushi interfaces both had a set speed. With the RSVP interface, participants were only shown one image at a time, and several commented that the image changed too fast. They stated that this sometimes led to mistakes, where they accidentally selected the image that followed their intended target.

With the Sushi selector, very different feedback was received. Some of the participants said that the images moved too slowly, and that they felt almost bored while completing the task. This occasionally made the participant feel frustrated when there were several False images in a row and they didn't need to respond. A few of the participants also said that they looked away or lost focus at these times, which could possibly have contributed to errors. Based on this feedback, a study was created to investigate the impact of changing the speed of the stream of images.

5.1 Study Task

In this study participants were once again shown a goal image containing a leaf, and were asked to categorize the test images according to whether they contained the same species of leaf. The dataset from the first study was reused in this study. This dataset consisted of three goal images, each with a corresponding set of 16 matching images, as well as 112 distracter images. There were a total of 160 images in the dataset. The order of the goal images was balanced using a Latin square.

The participants were shown blocks of 80 images, where each image was considered one trial. They were asked to categorize these images according to whether or not they matched the goal image. The method of categorizing the images was the same across the three interfaces, although the speed at which the images moved varied. Each of the goal images was shown for two blocks, or 160 trials. The same 160 images were used for each goal image, and in all conditions. The participants were required to complete six blocks (480 trials).

5.2 Design

The system from Study 1 was modified to create three speeds of the Sushi selector interface. Based on the feedback that the images moved too slowly, the Sushi selector from Study 1 was used as the Slow condition.

This study used a within-subjects design, with all participants completing the image categorization task using all three interface conditions. The independent variable was the speed at which the images moved across the screen. This study consisted of three conditions, which referred to the three speeds of the image stream. These conditions were: Slow; Medium; Fast. The conditions were balanced using a Latin square to account for possible sequencing effects.

The dependent variables were accuracy, perceived effort, and user preference. Accuracy was measured by counting the number of correct image categorizations. The participants categorized a total of 480 images in each condition, making 480 the maximum number of correct responses. The number of correct responses made in each condition was counted to calculate overall accuracy. Precision and recall were also measured as indicators of accuracy.

Perceived effort was measured using the version of the NASA-TLX [23] questionnaire that was created for the first study, which can be seen in Appendix B. This questionnaire breaks effort down into the following components: ease of use, ease of decision making, how overwhelming was the task, how rushed did the participant feel, the participant's subjective accuracy (whether or not they felt like their categorizations were correct), visual appeal, mental demand, and physical demand. The participants were asked to respond to statements that described aspects of these measures using a five point Likert scale from strongly disagree to strongly agree. The participants were asked to respond to the following statements:

- This style was easy to use,
- It was easy to make decisions about a specific image with this style,
- I felt overwhelmed by the number of choices with this style,
- I felt rushed while making decisions with this style,
- I was able to accurately identify matching images using this style,
- This interface was visually appealing,
- Using this selection style was mentally demanding, and
- Using this selection style was physically demanding.

User preference was measured by having the participants rank all seven of the interfaces from best to worst after they completed the tasks. This Summary questionnaire can be seen in Appendix C.2. The participants were asked which interface they thought allowed them to complete the task the fastest, and with which interface they thought their categorizations were the most accurate. This was done to compare the participants' perception of their completion time and accuracy to the measured results.

5.3 Study Conditions

This study consisted of three conditions, which were balanced between the participants using a Latin square rotation. Participants were required to select matching images from a dataset in all three conditions. The task required participants to categorize images of leaves according to whether or not they matched a specified Goal image.

Figure 5.1: Sushi selector interface.

All three conditions used versions of the Sushi selector interface from the first study. The Sushi selector presented a moving stream of test images underneath the Goal image. There were up to four images visible at one time, and the images moved across the screen from right to left, as shown in Figure 5.1. The participant would drag an image out of the moving stream in order to categorize it as a match. This can be seen in Figure 5.2. Non-matching images would be ignored, and allowed to advance off of the screen.

The **Slow** condition used the exact Sushi selector interface that was used in Study 1. The images moved

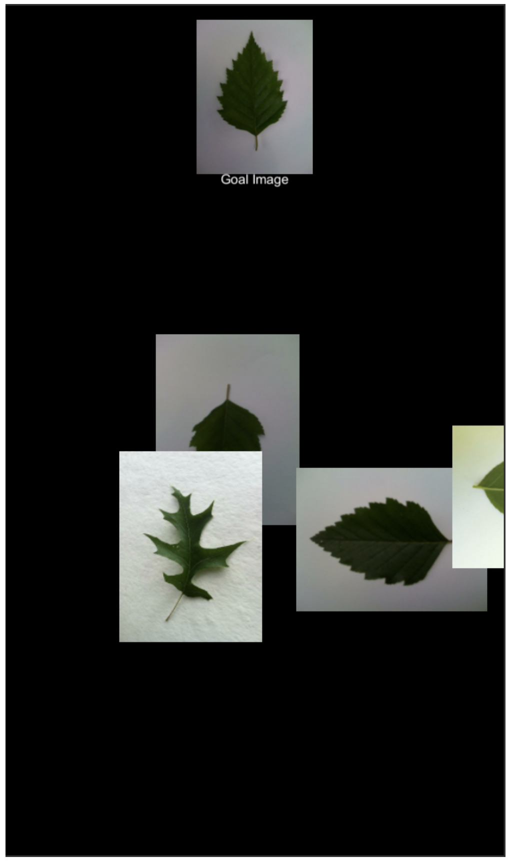

Figure 5.2: Sushi selector showing a "True" categorization.

at the same speed, taking approximately 5.0 seconds for each image to cross the screen. The speed of the image stream was increased to create the **Medium** condition. In this condition, each image was visible for approximately 3.9 seconds. The speed was increased once more to create the **Fast** condition. Each image was onscreen for approximately 2.7 seconds in this condition.

5.4 Procedure

After providing informed consent, the participants were told that they would be asked to complete an image search task with the images moving across the screen at three different speeds. They then filled out a demographics form detailing their sex, any colour vision deficiencies, and how often they use touchscreen devices like smartphones and tablets in their daily lives. This form can be seen in Appendix A.

After completing this form, they began the study with the first of the three image-stream speeds. They were given instructions on how to use the Sushi selector interface, and the action required to select matching images. They were then shown a training block consisting of 80 practice images in order to familiarize themselves with the speed of the interface. The practice images were pictures of dogs. The participant was asked to identify which dogs matched the breed of the goal image (which showed a husky). This was followed by 6 testing blocks each consisting of 80 images of leaves. They were allowed to take a break between blocks if they wished. The goal image was changed after the second and fourth blocks.

The Sushi selector from Study 1 was used in this study as the Slow condition. The Medium and Fast conditions used the same interface,but with the speed of the image stream increased slightly for each condition. The accuracy of the match or non-match labels that the participants provided for each image was measured in order to calculate precision and recall for each condition.

After completing the task, the participants were asked to complete the an effort questionnaire based on the NASA-TLX [23] in order to measure the usability of the interface at that speed. They then repeated this process for each of following speeds. Before each of the following tasks they were told that the interface worked in the same way, but that the images would move at a different speed. They didn't need to receive the further instructions again. They did complete the training block before each condition, to help them get used to the speed at which the images would move. They were asked to complete the TLX-style questionnaire again after each condition. After completing the TLX-style questionnaire for the final interface speed, the participants filled out a summary questionnaire asking them to rank the three conditions from best to worst, and to report in which condition they thought that they were able to complete the task while making the fewest errors.

5.5 Participants and Apparatus

This study was performed with 9 participants, all of whom were over the age of 18. Six of the participants were female, and three were male. None of the participants reported experiencing any colour vision deficiencies. Seven of the participants reported using touchscreen devices, such as smartphones or tablets, for more than 2 hours per day on average.

The study was conducted on a Microsoft Surface Pro 4 with a 12.3-inch display and a screen resolution of 2736 x 1824. The Surface was used as a tablet and kept in a portrait orientation. The software recorded all experimental data, including selection times and errors.

5.6 Data Analysis

Once again this study used the TLX-style questionnaire to measure the following metrics: easy to use, easy to make decisions, overwhelming, rushed, subjective accuracy, visual appeal, mental demand, and physical demand. The participants were asked to rate the items using a 5-point scale from "Strongly Disagree" to "Strongly Agree." The statements "It was easy to make decisions about a specific image with this style" and "I felt overwhelmed with the number of choices in this style." as well as the measures for subjective accuracy and visual appeal were again included in the questionnaire to check for differences between the different speeds of the image stream. There were spaces provided for the participants to elaborate on their responses or to provide additional feedback.

The system automatically recorded the number of correct and incorrect responses that the participants gave in each condition. This also allowed for comparison of the precision and recall data across all conditions.

RM-ANOVAs were run on the performance and questionnaire data, in order to determine which interfaces showed significant differences. Post-hoc comparisons were performed using the Tukey test. An alpha level of 0.05 was used for all statistical tests.

5.7 Results

5.7.1 Correct Responses

The ANOVA results for this study did not show significant differences for the total number of correct responses across the different image stream speeds at $p < .05$. Figure 5.3 shows the mean number of correct responses from the participants for each condition.

The data showed that the participants made the most correct categorizations when using the Slow interface, next was the Medium speed interface, and they made the fewest correct responses when using the Fast interface. With the Fast interface the participants made an average of 449.66 correct responses out of 480, with a standard deviation of 14.23 and a standard error of 4.74. Using the Medium interface the mean number of correct responses increased slightly to 453.77, with a standard deviation of 13.76 and a standard error of 4.59. In the Fast condition, the mean number of correct responses was 458.89 with a standard deviation of 9.21 and a standard error of 3.07. The diference between the Slow and Fast interfaces was not significant, with a p-value of .084.

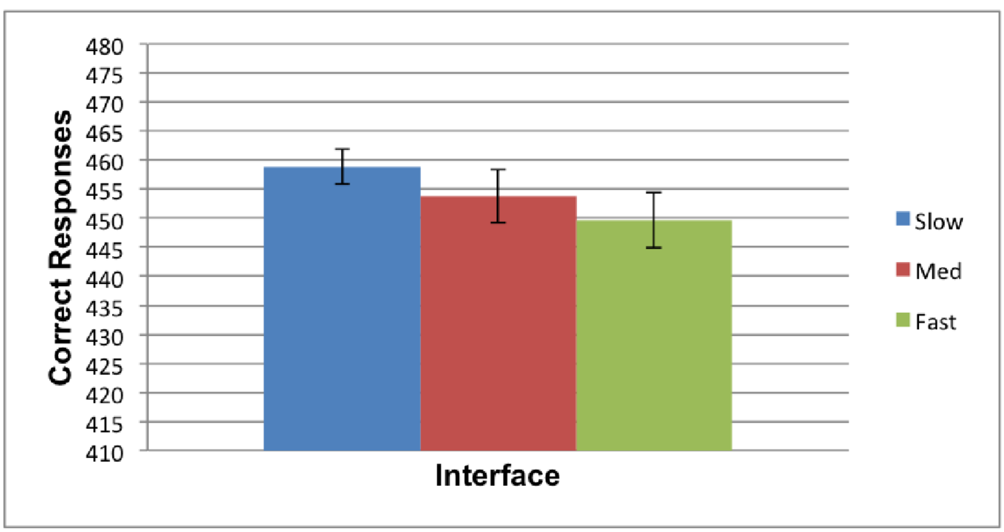

Figure 5.3: Mean number of correct responses by interface style. Error bars represent the standard error. The maximum possible number of correct responses was 480.

The results did not show significant differences in precision and recall across the three conditions. The mean precision values for the different interface styles can be seen in Figure 5.4. Once again there were slight variations, showing that the Slow interface had the highest precision rating and the Fast interface had the lowest, but these results were not significant. The participants had an average recall of .71 in the Fast condition, with a standard deviation of .18 and a standard error of .06. In the Medium condition the mean recall was .73, with a standard deviation of .18 and a standard error of .06. The Slow condition had the highest average recall at .77, with a standard deviation of .12 and a standard error of .04.

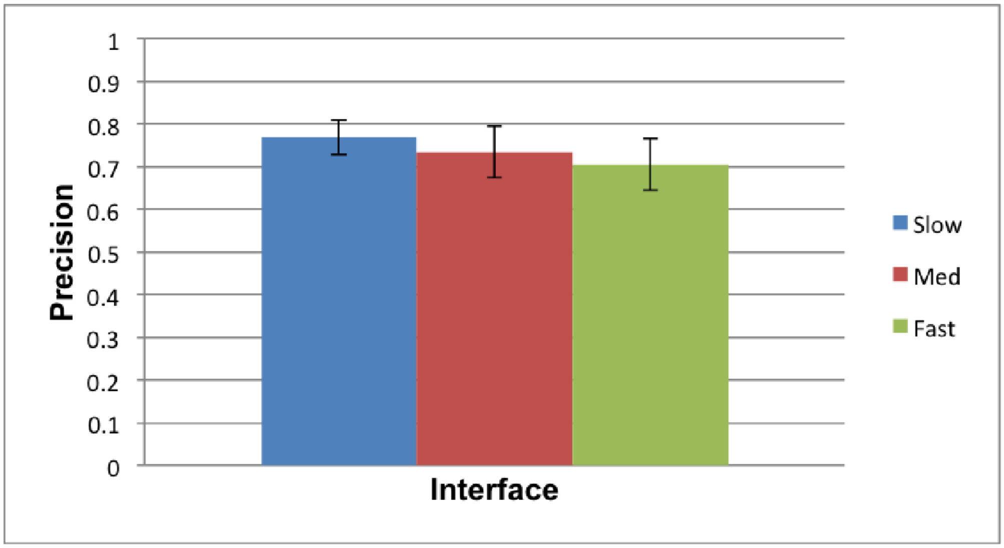

Figure 5.4: Mean precision by interface style. Error bars represent the standard error.

The mean recall values across the different interfaces can be seen in Figure 5.5. The Fast condition had an average participant recall of .78, with a standard deviation of .11 and a standard error of .04. In the Medium condition the mean recall was .83, with a standard deviation of .14 and a standard error of .05. The participants in the Slow condition had a mean recall rate of .84, with a standard deviation of .10 and a standard error of .03. Based on these values, the recall results also showed that the Fast interface had the lowest recall and the Slow interface had the highest, but these results were not significant.

5.7.2 Preference

After completing the image categorization task at all three speeds, participants were asked to rank the interfaces from 1 (best) to 3 (worst). The responses were reverse coded in order to graph the ratings more easily, which made 3 the highest possible score and 1 the lowest. The results of this question are shown in Figure 5.6. The Fast interface had an average rating of 1.33, with a standard deviation of .71 and a standard error of .24. The Medium interface had an average rating of 2.56, with a standard deviation of .53 and a standard error of .18. The Slow interface had an average rating of 2.11, with a standard deviation of .78 and

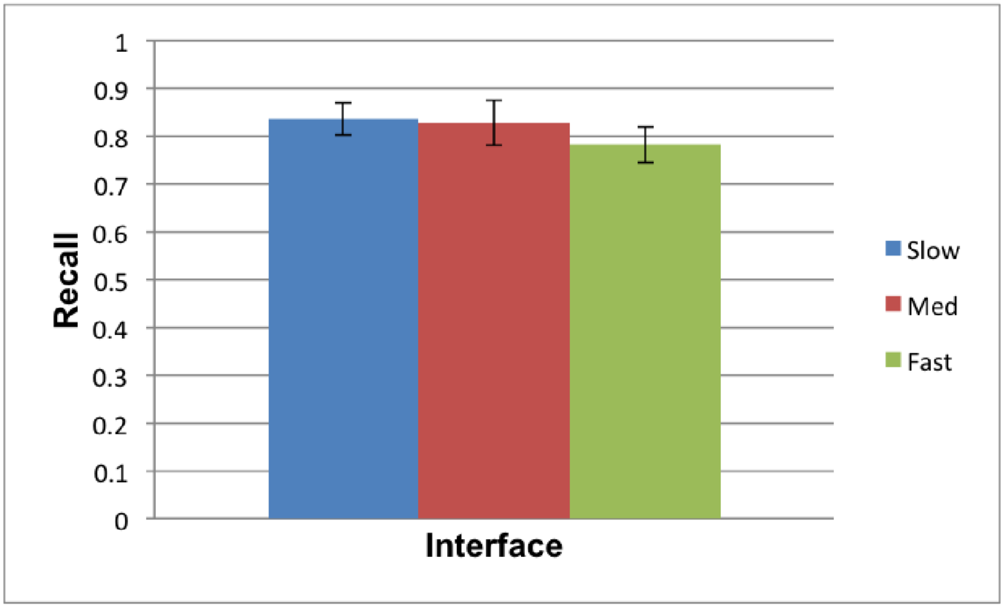

Figure 5.5: Mean recall by interface style. Error bars represent the standard error.

a standard error of .26. The Medium speed interface was the most preferred overall, with the Slow interface performing only slightly worse. The Fast interface was the most commonly disliked by the participants. The feedback received from participants stated that the images moved too quickly in the Fast condition, and that this made the task much more difficult.

5.7.3 Subjective Effort and Appeal

The TLX-style [23] questionnaire was used to measure the levels of mental and physical demand that using the interfaces had on the participants. This questionnaire also measured how overwhelmed and rushed the participants felt while using each interface. The TLX-style questionnaire results are shown in Figure 5.7.

The Slow interface was rated as significantly easier to use than the Fast interface, with a p-value of .021. The Medium interface was also rated as slightly easier to use than the Fast interface. This result was not significant, with a p-value of .152.

The participants found it significantly easier to make decisions about a specific image while using the Slow interface than with either the Medium or Fast interfaces. The p-values for this measure were .013 and 0, respectively. The Medium interface was rated as slightly easier to make decisions with than the Fast interface, but these results were not significant.

The participants felt significantly more rushed while using the Fast interface than with either of the other interfaces. The p-value for the Slow interface was 0, and for the Medium interface it was .001. The participants also felt more rushed while using the Medium interface than with the Slow interface, although this difference had an alpha value of .185 and was not significant.

The participants reported that they felt they were able to categorize the images more accurately while

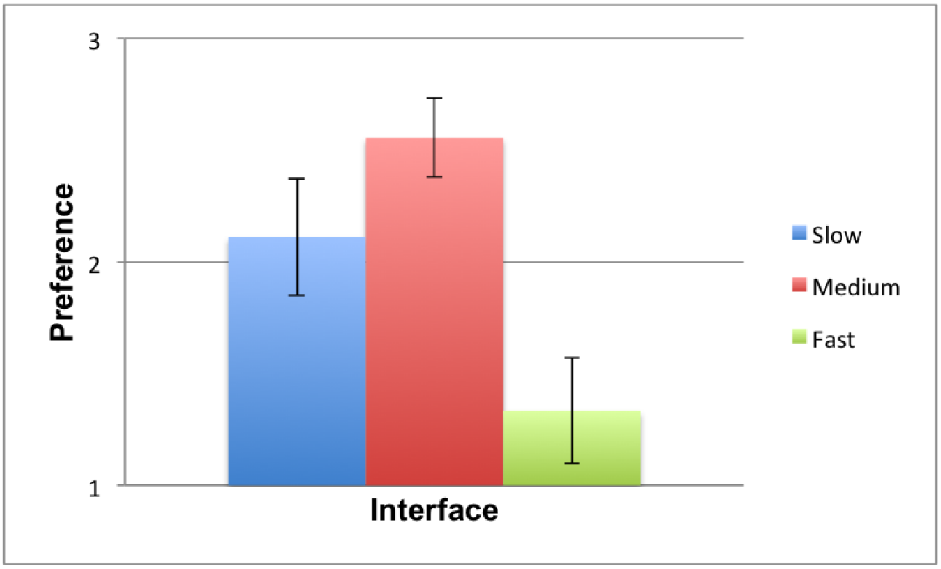

Figure 5.6: Average participant rankings of interfaces (reverse-coded; higher is better). Error bars represent the standard error.

using the Slow interface than the Fast interface, with a p-value of .009. They also reported feeling that they were slightly more accurate with the Medium interface than the Fast interface. This result was not significant, with a p-value of .131.

The participants reported feeling slightly more overwhelmed while using the Fast interface, but this result was not significant. There were no significant differences between the interfaces reported for the measures of visual appeal or physical demand. In both of these measures, the Slow interface performed slightly better than the other two with the Fast interface receiving the lowest ratings. This was also true for the measure of mental demand, although the difference between the Slow and Fast interface was not significant, with a p-value of .08.

5.7.4 Participant Comments

The participants were asked to provide comments or feedback at the end of each condition. The participants found the Slow interface fairly easy to use. Most of them felt that the interaction style was simple and intuitive. They also felt like they had enough time to accurately categorize the images in this condition. Some of the participants reported that the images moved too slowly, which occasionally led to boredom.

In the Medium speed condition, there were some mixed reviews. A few of the participants reported that the images moved too quickly, while others said that they preferred the slight speed increase.

Most of the participants felt that the images moved too quickly in the Fast condition. They felt rushed, and struggled to make decisions within the required time. A couple of the participants commented that they enjoyed the challenge of this interface, and felt like they were playing a game.

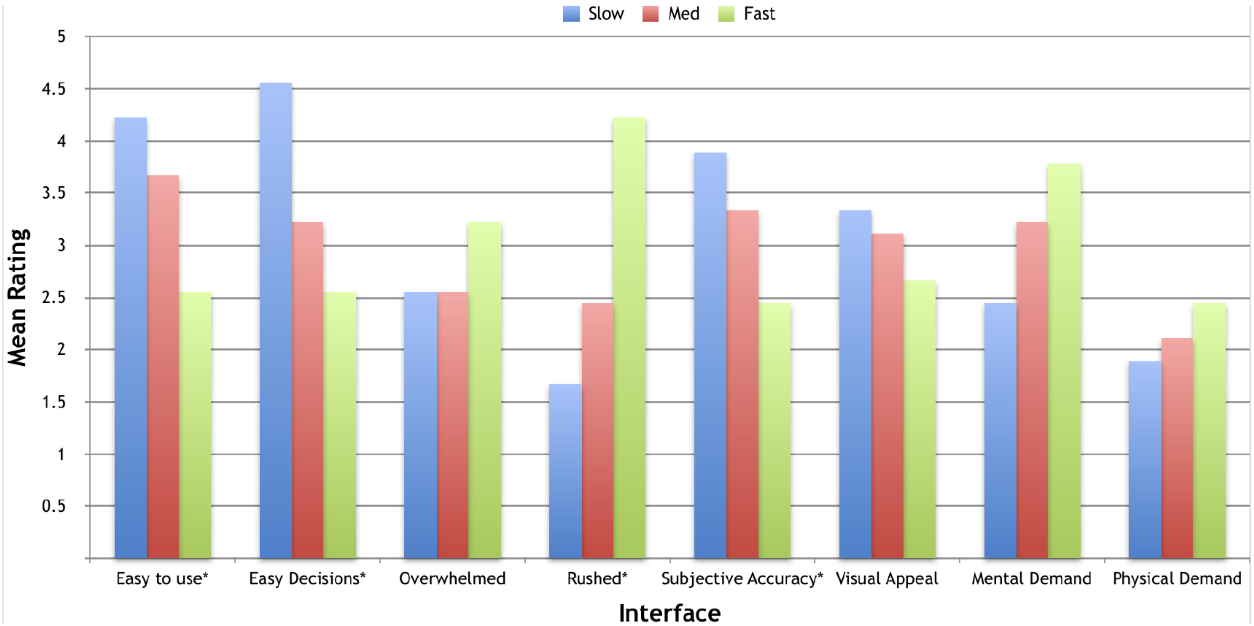

Figure 5.7: TLX-style questionnaire responses by interface style. Measures with significant results are starred.

5.8 Discussion

This investigation into the effect of the image stream speed on participant performance and experience provides five main results:

- The participants preferred the Medium and Slow interfaces over the Fast interface, which made them feel rushed;
- The Slow interface's ratings for ease of use were significantly higher than the Fast interface;
- It was significantly harder to make decisions with the Fast interface;
- The participants felt that they were less accurate when using the Fast interface;
- The Fast interface was slightly more mentally demanding, and caused the participants to feel overwhelmed.

The following section provides explanations for these findings.

5.8.1 Explanation and Interpretation of Results

Did the interface speed affect the accuracy?

The results did not show any significant differences in the overall accuracy between the different conditions. It was found that the responses were marginally more accurate when the participants were using the Slow interface than the Fast interface. There were no significant differences between the different speeds in regards to precision or recall.

There was some evidence of a speed-accuracy trade off. The participants were slightly more accurate with the Slow interface across all three measures of accuracy. The Fast interface performed the worst across these same measures, with the Medium interface falling somewhere in between. It is expected that if the study were to be run with more participants, these differences would become more pronounced.

The participants reported that they felt they were the least accurate in completing the task when they were using the Fast interface. The free-form responses that were received said that the images moved too quickly with the Fast image stream. They stated that they felt overwhelmed, and didn't always have enough time to accurately categorize the images before they left the screen.

Did the speed of the image stream affect the use of the interface?

Looking at the TLX-style [23] results shown in Figure 5.7, it can be seen that the Slow interface received the best rating across every single measure (high score for easy to use and make decisions, subjective accuracy, and visual appeal; low score for overwhelming, rushed, and mental and physical demand). The Fast interface had the worst rating, with the Medium interface falling somewhere in between. This was true across all of the measures except the overwhelmed measure, where the Medium and Slow interface had approximately the same result.

The participants gave the Fast interface the lowest preference rating of the three interfaces. They said that the images moved too fast, and this made the interface hard to use. They had trouble completing the task, and felt that they were less accurate when making decisions. This interface also took more effort to use, and left them feeling frustrated and overwhelmed. A couple of the participants did report this as their favourite interface. They said that the challenge involved in identifying the images quickly made using this interface fun. They said that needing to work harder to complete the task kept them interested, and that using this interface felt more like playing a game.

The Slow and Medium interfaces had similar results across most of the measures. The Slow interface did outperform the Medium interface in almost every measure, but most of these differences were not significant. The Medium interface received a slightly higher overall preference rating from the participants. The participants found that it was easier to make decisions about the pictures with the Slow interface, and felt that the Slow interface was slightly easier to use. Both the Slow and Medium interface performed well across the measures, and mixed feedback was received from the participants about which provided the best user experience.

Which interface speed was the best?

It was found that with every measure used, across both the TLX-style questionnaire and the logged experimental data, the Fast interface had the worst performance. The Slow interface performed the best, and the Medium interface was between the two. The participant feedback stated that the images moved too quickly with the Fast interface. This interface was also rated as more mentally demanding to use, and left the participants feeling rushed and overwhelmed. This was clearly the worst interface, in both usability and accuracy.

The Medium interface had the highest preference rating, and was closely followed by the Slow interface. These two interfaces received similar scores across all of the forms of measurement. There was mixed feedback as to which of these two interfaces was the overall winner. As was the case in study 1, some participants reported that the Slow interface was too slow, which was part of the motivation for this study. Some participants said that the image stream moved too slowly, and that this occasionally led to them losing some concentration or feeling almost bored. A few comments stated that the Medium interface moved slightly too fast, and that this made the task more difficult and frustrating. Based on this mixed feedback, it was determined that the ideal image stream speed for most participants would be somewhere between the speed of the Slow and Medium interfaces.

5.9 Summary

In the first study, it was discovered that the speed of the Sushi selector may have had an impact on the user experience and their performance. Feedback was received stating that the images moved too slowly, and some of the participants found this frustrating. In this chapter, a study was described in which three different image stream speeds were compared for the Sushi selector interface. A controlled study with nine participants compared the accuracy of the responses, and the usability of the interfaces across the three conditions. It was found that the Fast interface was clearly the worst across all of the measures. The Slow and Medium interfaces both performed well, and it is believed that the ideal image stream speed would be a compromise between these two.

Table 5.1: TLX-style ANOVA results for Study 2. Significant results are starred.

		Sum of Squares	df	Mean Square	F	Sig.
Easy to Use	Between Groups	12.963	2	6.481	4.348	0.024
	Within Groups	35.778	24	1.491		
	Total	48.741	26			
Easy Decisions	Between Groups	18.667	2	9.333	11.2	0
	Within Groups	20.000	24	0.833		
	Total	38.667	26			
Overwhelmed	Between Groups	2.667	2	1.333	1.143	0.336
	Within Groups	28.000	24	1.167		
	Total	30.667	26			
Rushed	Between Groups	30.889	2	15.444	18.742	0
	Within Groups	19.778	24	0.824		
	Total	50.667	26			
Subjective Accuracy	Between Groups	9.556	2	4.778	5.432	0.011
	Within Groups	21.111	24	0.880		
	Total	30.667	26			
Visual Appeal	Between Groups	2.074	2	1.037	1.087	0.353
	Within Groups	22.889	24	0.954		
	Total	24.963	26			
Mental Demand	Between Groups	8.074	2	4.037	2.595	0.095
	Within Groups	37.333	24	1.556		
	Total	45.407	26			
Physical Demand	Between Groups	1.407	2	0.704	0.422	0.66
	Within Groups	40.000	24	1.667		
	Total	41.407	26			

Chapter 6

Study 3 - Smartphone Classification

The first two studies both had participants use a tablet in order to complete the image categorization task. It was decided to use a portable device in order to test the interfaces on a common device that many people would be able to access. After completing the second study, it was time to see how the interfaces worked on an even more common, and smaller, device. It was decided to conduct a future study using a smartphone, instead of a tablet. For this study a Nexus 5 smartphone with a 4.95-inch display and a screen resolution of 1080 x 1920 was used. The phone was kept in a portrait orientation throughout the study.

There was interest in further work related to the impact of the complexity of the task that the participants were asked to complete. In the first study participants were asked to complete a fine-grained image classification task, where they were asked to match images of leaves according to their species. This is a specialized task, which could be difficult for the average user. People with little or no knowledge related to plant-science may not know what features to look for in leaf identification. Inaccurate responses may have been due to a lack of domain knowledge, instead of being related to the interface style.

In the third study, the impact of changing both the device used and the complexity of the task was examined. The task that was chosen for the participants to complete was based on some of the Captcha security checks that are seen on mobile websites. These tasks involve the participants looking at a selection of images and identifying all of the images that contain a common object like a street sign or a river.

6.1 Study Task

For this study, the task that the participants were asked to complete was changed. The participants were shown a Goal prompt asking them to select all images that contained a common and easily identifiable object. There were three goal prompts used in this study, each with a corresponding set of 16 matching images. The goal prompts were: "Select all images that contain a traffic light," "Select all images that contain a store front," and "Select all images that contain a street sign." The dataset contained an additional 112 distractor images, of street or city views containing none of the goal objects. The order of the goal images was balanced using a Latin square.

Similar to the earlier studies, participants were shown blocks of 80 images, with each image accounting for one trial. They were asked to categorize these images as matches or non-matches according to whether or not they contained the current goal object. The method of categorizing the images changed according to which interface style was in use at the time. Each goal image was visible for two blocks, or 160 trials, before changing. The same 160 images were shown for each goal objecte, and in all conditions. The task ended when the participant completed six blocks, or 480 trials.

6.2 Design

The system that had been used for the earlier two studies was modified to complete this experiment. Most of the same interfaces that were used in the first study were used again in the third. More details about this system can be found in Chapter 3.

This study used a within-subjects design, so all of the participants were asked to complete the image categorization task using all six interfaces. The independent variable was the type of interface used. This study consisted of six conditions, referring to the interaction style of each of the interfaces. These conditions were: Small Grid (4 x 5); Large Grid (5 x 8); Swipe; Preview Swipe; Sushi selector; RSVP. The conditions were balanced using a Latin square to account for possible sequencing effects.

The dependent variables in this study were completion time, accuracy, perceived effort, and participant preference. Completion time was automatically recorded by the system timer. The timer did not increase between conditions, or if the participant paused between blocks of images. Counting started when the participant pressed the button to begin the trials for a block, and stopped when the last test image in that block left the screen.

Accuracy referred to the overall accuracy of the participant's categorization labels. The total number of images each participant was asked to categorize using each interface was 480, making this the maximum possible number of correct responses. Total number of correct categorizations made using each interface was counted, as the primary measure of accuracy. Precision and recall values were also used as indicators of accuracy.

Perceived effort was measured using the version of the NASA-TLX [23] questionnaire, created for the earlier studies and shown in Appendix B. This questionnaire breaks effort down into the following components: ease of use, ease of decision making, how overwhelming was the task, how rushed did the participant feel, the participant's subjective accuracy (whether or not they felt like their categorizations were correct), visual appeal, mental demand, and physical demand. The participants were asked to respond to statements that described aspects of these measures using a five point Likert scale from strongly disagree to strongly agree. The participants were asked to respond to the following statements:

- This style was easy to use,
- It was easy to make decisions about a specific image with this style,
- I felt overwhelmed by the number of choices with this style,

- I felt rushed while making decisions with this style,
- I was able to accurately identify matching images using this style,
- This interface was visually appealing,
- Using this selection style was mentally demanding, and
- Using this selection style was physically demanding.

Participant preference was measured by having the participants rank the six interfaces from best to worst at the end of the study. This Summary questionnaire can be seen in Appendix C.3. Participants were asked to identify which interface they thought allowed them to complete the task the most quickly, and with which interface they thought they were able to make the most accurate categorizations. These results were used to compare the participants' perception of their completion time and accuracy to the measured results.

6.3 Study Conditions

This study consisted of six conditions, which were balanced between the participants using a Latin square rotation. In all of the conditions, participants were asked to categorize a set of test images according to whether or not they contained a specified goal object.

In the **Small Grid (4 x 5)** condition, the participants were shown the Goal prompt above a 4x5 grid containing 20 test images. They were able to select matching images by tapping on them. Selected images were displayed with a green border, in order to allow participants to keep track of their selections. This interface is shown in Figure 6.1. Tapping on an already selected image would deselect it. This allowed participants to correct potential mistakes.

When the participant decided that they had selected all of the matching images on a page they could push the "next" button at the right side of the page to advance. Pushing this button on the last page of images in a block, this button would reveal a screen informing participants that they were allowed to pause and take a short break if they wished. When the participant was prepared to continue, they would push the button on this page to resume the task. After the final block of images, pushing the "next" button brought the participant to a page asking them to stop and answer questions. Both of these screens can be seen in Figure 6.2.

The largest grid size (8x10) was eliminated for this study, since the images had already been reported as being too small on the tablet, and this problem would be made worse by using a smartphone with a smaller screen. The Medium Grid (5 x 8) interface from Study 1 became the largest grid size, and is referred to as the Large Grid (5 x 8) for this study. The Zoom Grid interface was also removed, since the effects of this feature had already been explored.

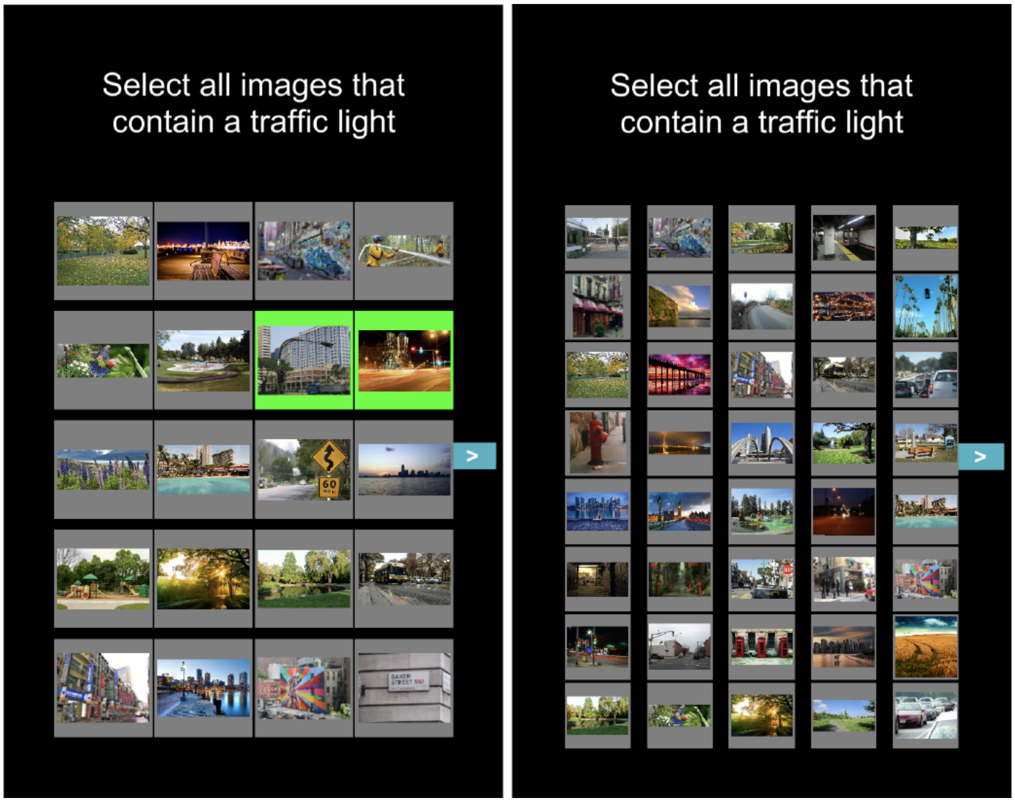

Figure 6.1: Small Grid (4 x 5) with images selected (left), Large Grid (5 x 8)(right)

The **Large Grid (5 x 8)** contained 40 images per grid page (5x8 images), and is shown in Figure 6.1. The additional images meant that each image had to be smaller in order to fit on the screen. These images were even smaller than in previous studies, due to the decreased screen size of the mobile device. Once again, the participants were asked to select matching images by tapping on them, and proceeded to the next grid page once they had selected all of matches on a page.

The **Rapid Serial Visual Presentation (RSVP)** interface showed one test image at a time. This image was displayed underneath the current Goal prompt. Each test image was shown for one second, before automatically advancing to the next image. Tapping on the visible image would select it, or categorize it as a match. Selected images would turn green to indicate the selection. This interface can be seen in Figure 6.3, with an example of a selection. Non-matching images would remain unselected. After 80 images had been displayed, completing an image block, a button would appear and the participant would be prompted to press it in order to continue the task. At the end of the final block, the prompt would tell the participants to stop and answer questions before continuing.

The **Swipe** interface showed the participants one image at a time, positioned underneath the Goal prompt. Touching the test image and dragging it to the left (swiping left) would categorize it as a non-matching image. Swiping the test image to the right would categorize it as a match. The sides of the screen were labeled, in order to remind the participants which direction to swipe for each action. A red 'X' was

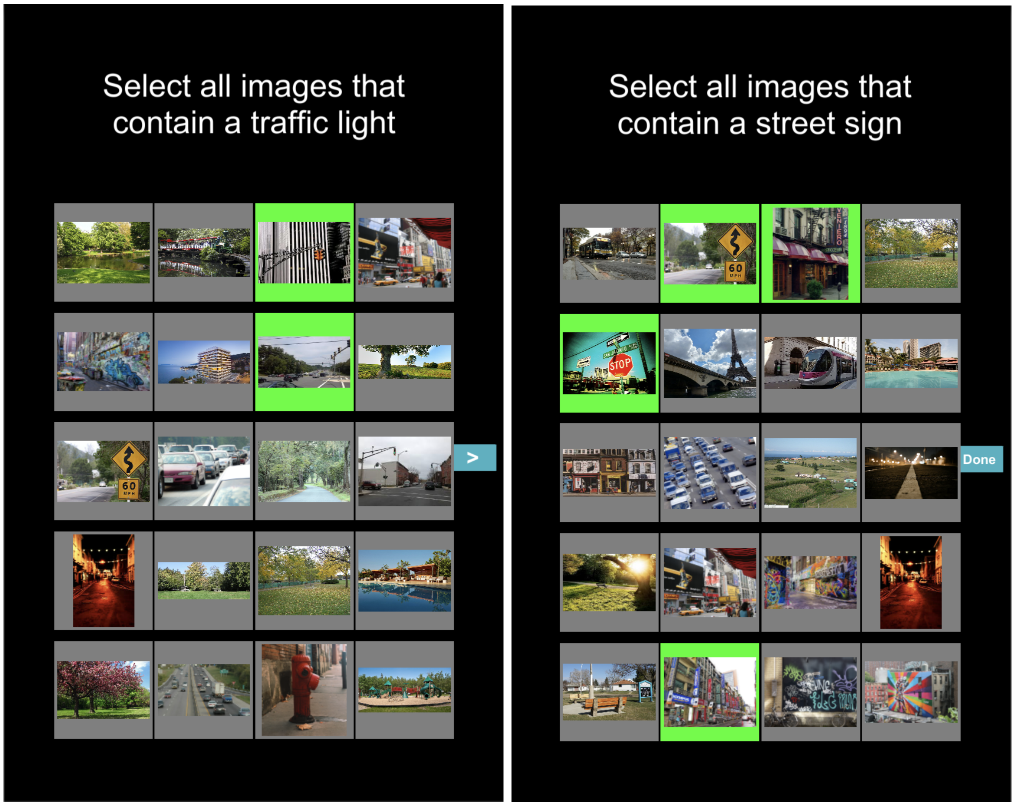

Figure 6.2: Grid interface showing the "Next" and "Done" buttons.

shown on the left side of the test image, and a green 'check' icon was on the right. While dragging the test image to the right a green overlay indicated a matching categorization. A red overlay when swiping to the left indicated a non-match. After the participant had categorized all images in a block, a button would appear with a prompt informing them that they could take a break and continue when they were ready. After the final block of images, the prompt informed the participants that it was time to answer questions. This interface can be seen in Figure 6.4.

A Preview feature was added to the Swipe interface, to see if this improved the use of the interface. The **Preview Swipe** interface worked in a similar manner as the Swipe interface. The participants were shown an image, and asked to swipe left if the image didn't match the prompt, and to swipe right if it did match. The difference was that with the Preview feature, the participants could also see the next two images that they would be asked to categorize. The next image was shown directly above the current image, with a reduced alpha value to make it appear slightly greyed out. The image following the next image was shown above the next image, with the alpha value further reduced. When the participant swiped the current image, the other two images that were on the screen would move down into their new positions (next to current, third image to next). It was expected that the Preview feature would allow the participants to make decisions more quickly, since they would be able to see each image prior to being asked to categorize it. This interface is shown in Figure 6.5.

Figure 6.3: RSVP interface (left) RSVP with a selected image (right).

The **Sushi** selector interface showed the participants a moving stream of test images underneath the Goal image. There were up to four images visible at one time. Study 2 showed that the ideal speed for the Sushi selector interface fell between the Slow and Medium speeds. In the Slow condition the images had been onscreen for 5.0 seconds, while in the Medium condition they were onscreen for 3.9 seconds. The Sushi selector in this third study displayed the images for approximately 4.5 seconds. This speed was chosen because it was approximately halfway in between the two preferred speeds from the earlier study. The images moved across the screen from right to left automatically. In order to select a matching image, the participant would drag it out of the moving stream. This is shown in Figure 6.6. Non-matching images were left unselected, and allowed to advance off of the screen.

The Sushi selector, Preview Swipe, and both Grid interfaces all allowed participants to compare features across multiple images at once. The RSVP and Swipe interfaces only showed one individual test image at a time, not allowing direct comparison between multiple test images.

The RSVP and Sushi selector interfaces both had speed controlled by the system. The images automatically advanced, without input from the participant. The Swipe, Preview Swipe, and both Grid interfaces allowed the participants to set the pace of the task. They were able to control the speed at which the images advanced, and could move through the dataset making selections at their own pace.

The Swipe and Preview Swipe interfaces required the participants to take an explicit action to categorize

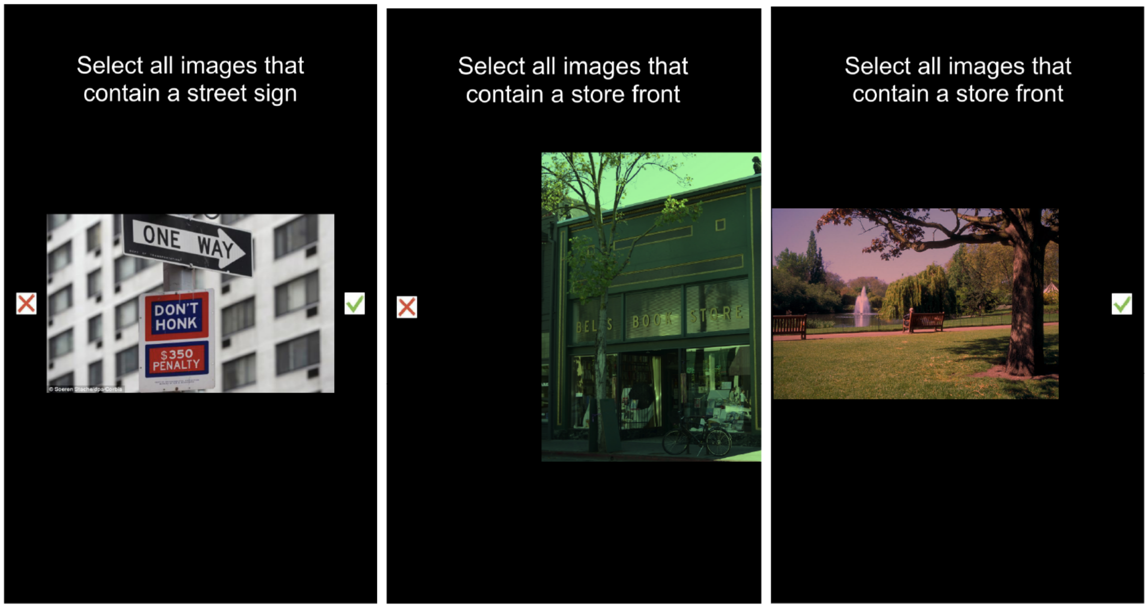

Figure 6.4: Swipe interface(left), with a True selection (centre), with a false selection (right).

non-matching images, as well as matches. Swiping to the right categorized an image as a match, while swiping to the left indicated a non-match. The RSVP, Sushi selector, and both Grid interfaces only required an action to select matching images. Non-matching images were left deselected by default, and did not require any action from the participant.

6.4 Procedure

After providing informed consent, it was explained to the participants were that they would be asked to complete an image categorization task using six variations of interface designs. They were then asked to fill out a demographics form detailing their sex, any colour vision deficiencies they might have, and how often they use touchscreen devices, such as smartphones and tablets, in their daily lives. This form can be seen in Appendix A.

Following this form, they began the study with the first of the six interfaces. The participants were given instructions on how to use the interface and select matching images. They then completed a training block consisting of 80 practice images, in order to become familiar with the interface. The practice images all contained dogs, and the prompt asked the participants to select all of the huskies. This was followed by 6 testing blocks each consisting of 80 images. They had the opportunity to take a break between blocks if they wished. The goal prompt changed after the second and fourth blocks.

In the Small Grid (4 x 5) interface, there were 20 images per page and the participants flipped through four pages of images for each of the six blocks. The Large Grid (5 x 8) contained 40 images per page, so the participants flipped through two pages of images per block. In the Sushi selector, Swipe, RSVP, and

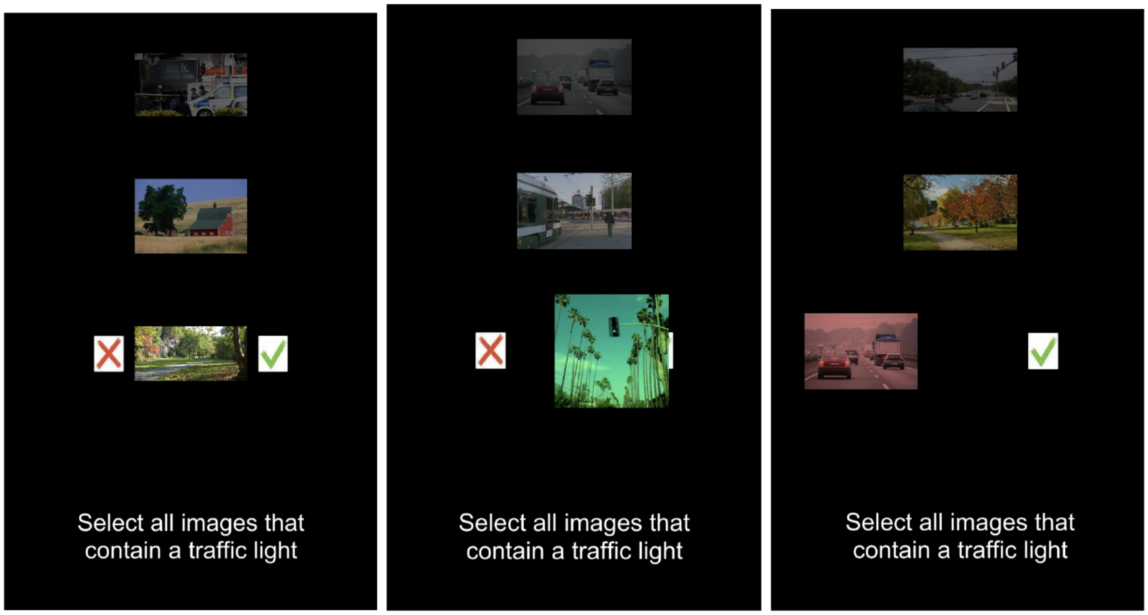

Figure 6.5: Preview Swipe interface(left), with a True selection (centre), with a false selection (right).

Preview Swipe interfaces, the participants saw the images in six blocks of 80 images, with the option to take a short break between blocks.

The participants were timed while they completed the task, to provide information on which selection style was the fastest or most efficient. The accuracy of the match or non-match labels that the participants provided for each image was also measured in order to calculate precision and recall for each condition.

After finishing the task with each interface, the participants were asked to complete the NASA-TLX [23] questionnaire. They repeated this process with each of the next five interfaces. After completing the TLX-style questionnaire for the sixth and final interface, participants were asked to fill out a summary questionnaire asking them to rank the six interfaces from best to worst, and to report which interface allowed them to complete the task the most quickly, and with which interface they thought that they made the fewest errors.

6.5 Participants and Apparatus

This study was performed with 18 participants, all of whom were over the age of 18. There were 8 male and 10 female participants. One of the participants reported having colour vision deficiencies with regards to red and green. Seventeen of the participants reported using touchscreen devices, such as smartphones or tablets, for at least one hour per day, with thirteen of those participants reporting the use of the devices for more than two hours per day on average.

The study was conducted on a Nexus 5 smartphone which had a 4.95-inch display and a screen resolution of 1080 x 1920. The phone was kept in a portrait orientation throughout the study. The software recorded

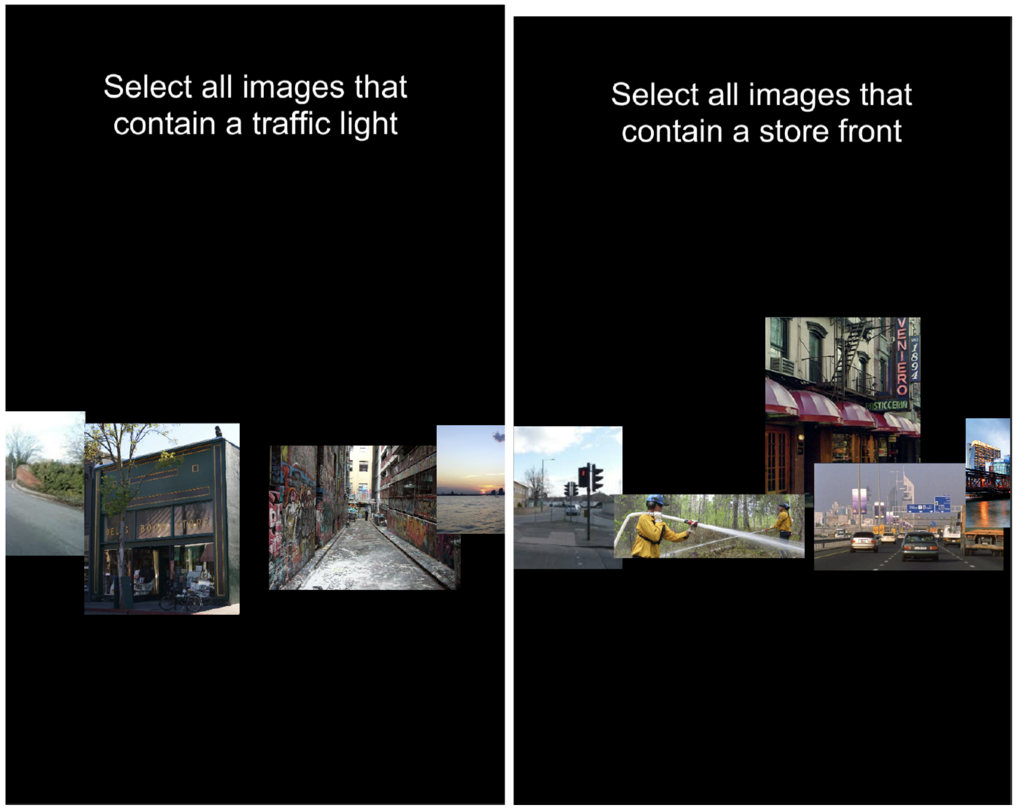

Figure 6.6: Sushi selector interface (left), with a selected image (right).

all experimental data, including selection times and errors.

6.6 Data Analysis

The TLX-style questionnaire was used to measure the same metrics as the earlier studies, which included: easy to use, easy to make decisions, overwhelming, rushed, subjective accuracy, visual appeal, mental demand, and physical demand. The questionnaire items were rated using a 5-point scale from "Strongly Disagree" to "Strongly Agree." Additional spaces were provided where the participants could leave free-form comments and feedback about the different interface styles.

The system recorded all participant responses in order to measure the accuracy of their classifications. This also allowed for comparison of precision and recall values across the conditions.

RM-ANOVAs were run on all of the experimental data, in order compare the various interaction techniques. Post-hoc comparisons were performed using the Tukey test. An alpha level of 0.05 was used for all statistical tests.

Table 6.1: Time ANOVA results for Study 3. Significant results are starred.

		Sum of Squares	df	Mean Square	F	Sig.
Time	Between Groups	1307871.951	5	261574.39	24.606	0*
	Within Groups	1084303.496	102	10630.426		
	Total	2392175.448	107			

6.7 Results

6.7.1 Completion Time

Figure 6.7 shows the mean completion time for each of the interface styles. The ANOVA results can be seen in Table 6.1. It can easily be seen that there are significant differences across the different interaction techniques. The full post-hoc comparisons (performed using the Tukey test) can be seen in Appendix I.

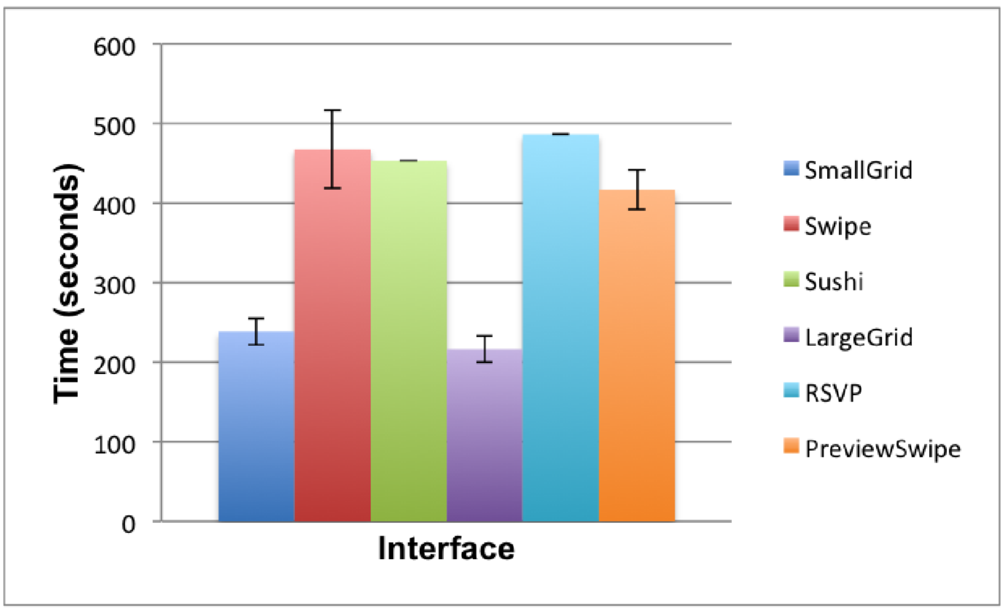

Figure 6.7: Mean task completion time by interface style. Error bars represent the standard error.

The Fastest condition was the Large Grid (5 x 8), with an average completion time of 216.19 seconds. The standard deviation was 72.99 and the standard error was 17.20. While using the Small Grid (4 x 5) interface, the participants had an average time of 238.15 seconds, with a standard deviation of 70.17 and a standard error of 16.54. The Preview Swipe condition took an average of 416.94 seconds, with a standard deviation of 105.11 and a standard error of 24.78. The mean completion time in the Sushi condition was 453.33 seconds, with a standard deviation of .60 and a standard error of .14. In the Swipe condition the mean completion time was 467.46 seconds with a standard deviation of 206.11 and a standard error of 48.58. The

RSVP condition took the longest to complete, with a mean time of 486.36 seconds. The standard deviation was .39 and the standard error was .09.

The two grid variations (Small, and Large) allowed the participants to complete the task significantly faster than the other four interfaces. Follow-up t-tests showed that both Grid interfaces were significantly faster than the other four interfaces ($p < 0.0001$). There were no significant differences between the Small Grid (4 x 5) and the Large Grid (5 x 8), or among the other four conditions.

Table 6.2: Responses to the questions "With which interface were you the fastest?" and "With which interface did you make the fewest errors?" The numbers show how many participants selected each interface as their response.

Interface	With which interface were you the fastest?	With which interface did you make the fewest errors?
Small Grid (4 x 5)	1	3
Swipe	2	4
Sushi	3	4
Large Grid (5 x 8)	1	1
RSVP	2	2
Preview Swipe	8	4

The participants were asked with which interface they felt that they had completed the task most quickly. These responses are shown in Table 6.2. The most common response from the participants was the Preview Swipe interface, with eight responses. Only two participants felt that either of the Grid interfaces allowed them to complete the task the most quickly, compared to ten participants that selected one of the Swipe interfaces. This is in spite of the fact that the timing results show that the Grid interfaces took around half as long as the Swipe interfaces.

6.7.2 Correct Responses

Figure 6.8 shows the mean number of overall correct responses from the participants for each condition. There could be a maximum of 480 correct responses. The ANOVA data did not show any significant diference in the number of correct categorizations across the conditions. The RSVP interface had the highest average at 463.67, with a standard deviation of 37.98 and a standard error of 8.95. The mean number of correct responses with the Preview Swipe interface was 458.72, with a standard deviation of 43.63 and a standard error of 10.28. The Swipe interface had a very similar average at 458.22, with a standard deviation of 11.92 and a standard error of 2.81. Using the Small Grid (4 x 5) interface, the mean number of correct responses was 456.22, with a standard deviation of 12.14 and a standard error of 2.86. The Sushi selector had a mean of 455.44 with a standard deviation of 17.00 and a standard error of 4.01. The Large Grid (5 x 8) condition had the lowest average at 453.11, with a standard deviation of 12.98 and a standard error of 3.06.

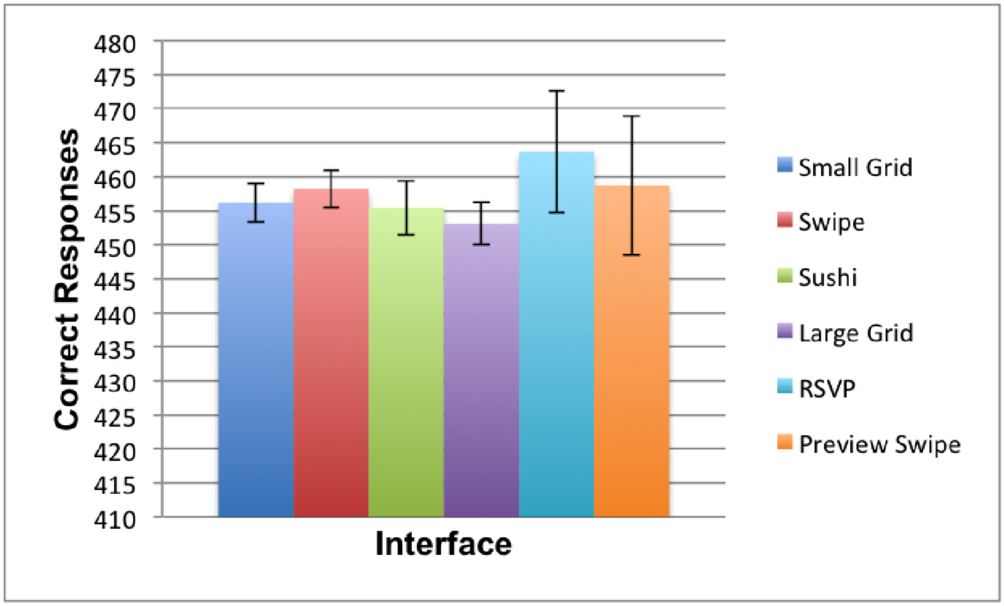

Figure 6.8: Mean number of correct responses by interface style. Error bars represent the standard error. The maximum possible number of correct responses was 480.

The results did not show any significant differences in precision across the different interfaces. Figure 6.9 shows the mean precision values across the conditions. The Large Grid (5 x 8) had the highest precision value, at .83. The standard deviation was .20 and the standard error was .05. The Small Grid (4 x 5) and Swipe interfaces both had a precision result of .80. The Small Grid (4 x 5) had a standard deviation of .17 and a standard error of .04, while the Swipe interface had a standard deviation of .15 and a standard error of .04. The mean precision in the Sushi condition was .77 with a standard deviation of .18 and a standard error of .04. The RSVP interface had a mean of .75, with a standard deviation of .16 and a standard error of .04. The lowest average was the Preview Swipe condition at .73 with a standard deviation of .21, and a standard error of .05. The lowest p-value across all of the comparisons was between the Large Grid (5 x 8) and the Preview Swipe interfaces, where $p = .585$.

Figure 6.10 shows the mean recall values across the conditions. The highest mean recall value was the Sushi selector at .81, with a standard deviation of .08 and a standard error of .02. The RSVP interface had a mean of .79 with a standard deviation of .10 and a standard error of .02. The Preview Swipe interface was only slightly lower, at .78 with a standard deviation of .13 and a standard error of .03. The Swipe condition was also close, with a mean recall of .77. The standard deviation was .13, while the standard error was .03. The Small Grid (4 x 5) condition had a mean recall of .72 with a standard deviation of .09 and a standard error of .02. The lowest recall value was the Large Grid (5 x 8) condition at .61, with a standard deviation of .12 and a standard error of .03.

The Large Grid (5 x 8) had significantly lower recall than all of the other interface styles. The p-value was .023 when compared to the Small Grid (4 x 5), and 0 when compared to the other four interfaces. The

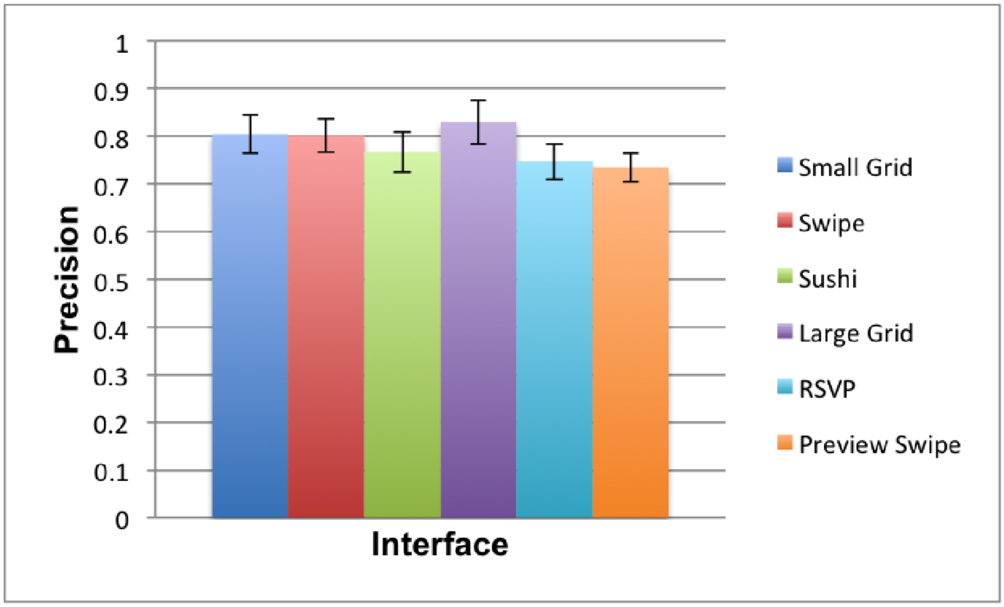

Figure 6.9: Mean precision by interface style. Error bars represent the standard error.

Sushi selector showed the highest recall, although this was not significant ($p = .156$) when compared to the Small Grid (4 x 5) and even less significant for the other four interfaces.

6.7.3 Preference

At the end of the study, the participants were asked to rank all of the interfaces that they used from best to worst in order of preference. The responses were reverse coded in order to more easily graph the interface ratings. This made 6 the highest possible score and 1 the lowest. Figure 6.11 shows the mean rankings that the interfaces received. It was found that the two Swipe-based interfaces (Swipe and Preview Swipe) were strongly preferred over the other four interfaces.

The Swipe interface had the highest score at 4.67, with a standard deviation of 1.24 and a standard error of .29. The Preview Swipe interface was a close second at 4.50, with a standard deviation of 1.5 and a standard error of .36. The Large Grid (5 x 8) had the third highest score at 3.11, with a standard deviation of 1.41 and a standard error of .33. This was closely followed by the Small Grid (4 x 5) at 3.06, with a standard deviation of 1.47 and a standard error of .35. The Sushi condition had a mean rating of 2.89 with a standard deviation of 1.88 and a standard error of .33. The RSVP interface scored slightly lower at 2.83, with a standard deviation of 1.76 and a standard error of .41.

6.7.4 Subjective Effort and Appeal

A version of the NASA-TLX [23] questionnaire was used to measure the levels of mental and physical demand that completing the image categorization task with the different interfaces had on the participants. This

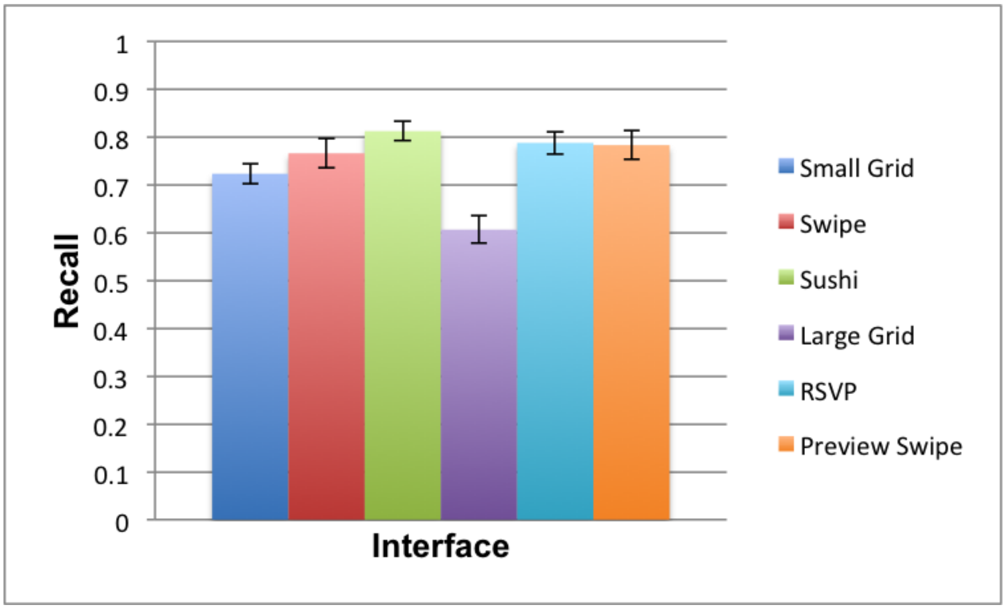

Figure 6.10: Mean recall by interface style. Error bars represent the standard error.

questionnaire was also used to measure how overwhelmed and rushed the participants felt while completing the task. The TLX-style results are shown in Figure 6.12. Full ANOVA results for the TLX-style questionnaire can be seen in Appendix H. Post-hoc comparisons (performed using the Tukey test) can be seen in Appendix I.

The Small Grid (4 x 5) was significantly easier to use than the Large Grid (5 x 8), with a p-value of .025. The Large Grid (5 x 8) interface was also significantly harder to use than the Swipe or Preview Swipe interfaces, with p-values of 0 and .001 respectively. The Swipe interface was rated as the easiest to use. It was also significantly easier to use than the Sushi selector or RSVP interface, with p-values of .007 for both. There was no significant difference between the Swipe and Preview Swipe interfaces. The difference between the Preview Swipe interface and the RSVP and Sushi interfaces was not significant, with a p-value of .104 for both.

The participants also found it to be easier to make decisions with the two Swipe interfaces than with the other interaction styles. The participants found it significantly easier to make decisions about a specific image with the Swipe interface than with either the RSVP or Large Grid (5 x 8) interfaces. It was also significantly easier to make decisions with the Preview Swipe interface than with the Large Grid (5 x 8) interface. The difference between the RSVP and Preview Swipe interfaces was not significant, with $p = .069$.

The participants found the Swipe interface significantly less overwhelming than the Sushi, Large Grid (5 x 8), and RSVP interfaces. The difference between the Small Grid (4 x 5) and the RSVP interface was not significant, with $p = .098$. There were no other significant differences among the interfaces.

Participants felt the most rushed while using the RSVP interface. This difference was significant when compared to all of the other interfaces, except for the Sushi selector. They felt significantly more rushed

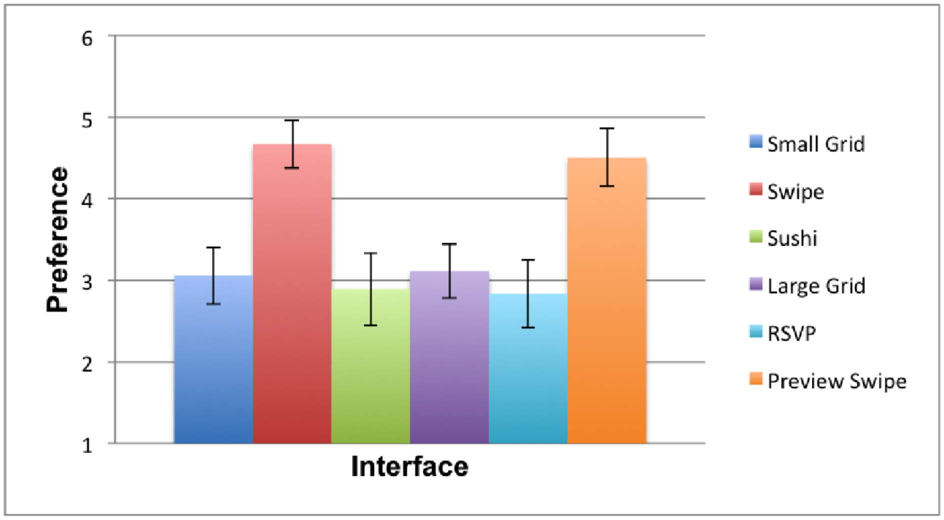

Figure 6.11: Average participant rankings of interfaces (reverse-coded; higher is better). Error bars represent the standard error.

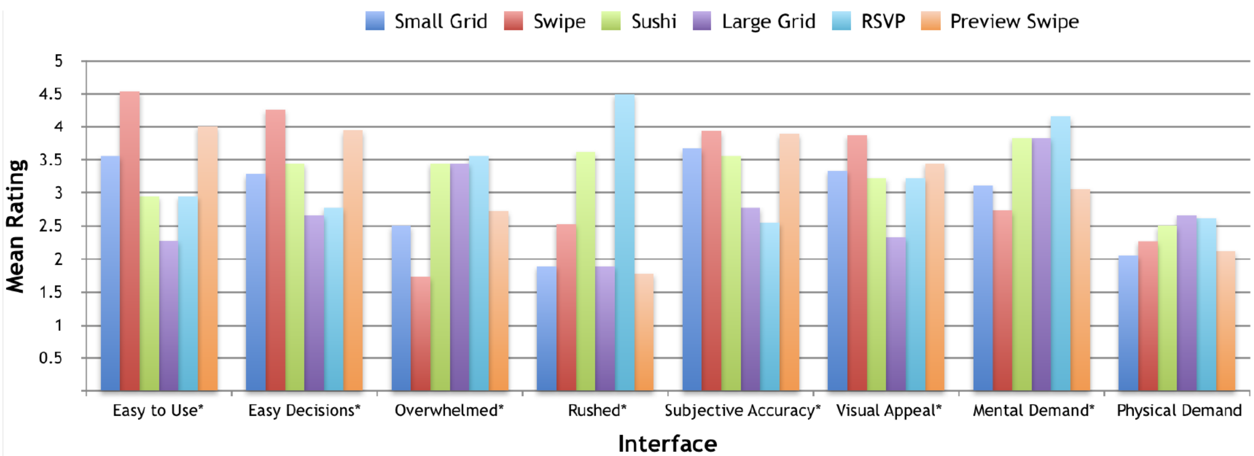

Figure 6.12: TLX-style questionnaire responses by interface style. Measures with significant results are starred.

with the Sushi selector than with the Preview Swipe interface and both of the Grid interfaces. The diference between the Sushi selector and the Swipe interface was not significant, with p = .097.

The participants felt that their responses were significantly more accurate when using the Swipe interface than either the Large Grid (5 x 8) or RSVP interfaces. The Preview Swipe interface scored significantly higher than the RSVP interface, although the difference between it and the Large Grid (5 x 8) was not significant with p = .055. There were no other significant differences.

The Swipe interface was rated as the most visually appealing. This difference was significant when compared to the Large Grid (5 x 8). The differences between the Large Grid (5 x 8) and the Small Grid (4 x 5) and Preview Swipe interfaces were the most pronounced, with p-values of .123 and .063. There were no other significant differences.

The RSVP interface was the most mentally demanding to use. It was significantly more demanding than the Swipe interface, with a marginal difference between it and the Preview Swipe ($p = .056$) and Small Grid (4 x 5) ($p = .08$) interfaces. The Swipe interface was also significantly less mentally demanding to use than either the Large Grid (5 x 8) or Sushi selector interfaces. There were no significant differences across the conditions with regards to physical demand.

6.8 Discussion

This study provided five main results:

- The two grid-based interfaces allowed participants to complete the image selection task significantly faster than the non-grid interfaces;
- There were significant effects on recall across interfaces;
- The participants preferred the two Swipe-based interfaces over the other interaction styles;
- There were significant differences in how easy it was to use or make decisions with the different interface styles;
- The participants felt more rushed while using the Sushi and RSVP interfaces, and the Swipe interface made them feel the least overwhelmed;

6.8.1 Explanation and Interpretation of Results

Were the grid-based interfaces still faster than the non-grid interfaces?

This study confirmed the earlier result, which showed that the grid-based interfaces were much faster than the non-grid interfaces. The participants reported feeling more rushed when using the RSVP and Sushi selector interfaces than any of the other four. These two interfaces were the only ones where the images changed or moved automatically. With the RSVP interface an image was displayed for one second before switching to the next image, while with the Sushi selector the images moved in a constant stream and were each visible for about 4.5 seconds. Since the participants were not able to set their own speed, some reported feeling rushed or overwhelmed.

The Sushi selector, Preview Swipe, and Grid interfaces allowed the participants to view multiple images at the same time. The Swipe and RSVP images only showed a single image at any given time. Results showed that the participants were slightly faster with the Preview Swipe interface than the regular Swipe interface, but this difference was not significant. As in the earlier study, participants commented that having multiple images available allowed them to compare the images to each other. This also allowed the participants to scan over or process multiple images simultaneously. The participant was able to begin making categorization decisions about another image (or multiple other images) while performing the action required to categorize a specific image. This decreased the time necessary to make a categorization decision about each individual image on average. Since the decision and categorization actions were able to take place simultaneously for multiple images, the participant was once again able to complete the task more quickly overall. This confirms the timing results from the first study.

The two Swipe-based interfaces were once again set apart since they require the participant to take an explicit action to categorize every single image as either True or False. The other four interface styles would interpret no action as a negative classification, while in the Swipe-based interfaces this required a left-swipe motion. The participants reported that the Swipe interfaces were the easiest to use, and allowed them to make decisions the most easily. Ten out the eighteen participants (more than half) also responded that they felt like one of the Swipe-based interfaces allowed them to complete the task the most quickly, compared to only two who responded with one of the Grid interfaces. In spite of this, results showed that it took almost twice as long to complete the task with the two Swipe interfaces than with either of the Grid interfaces. The extra time required to categorize each False image added up to a significant amount of time overall.

Did the selection style affect accuracy?

There were no significant differences in the overall accuracy of the categorization results across the different conditions. The data showed slightly less accurate responses with the Large Grid (5 x 8) interface, and slightly more accuracy with the RSVP interface, but neither of these effects were significant. Precision and recall were also measured as tests of accuracy. There were no significant results in the precision data across conditions.

The results did show that recall was significantly lower with the Large Grid (5 x 8) interface than with any of the other interfaces. The participants complained that the images on the screen were too small with this interface, and that they were hard to see. Not being able to see the objects in the image clearly led to fewer overall selections. If the images were too small to see the traffic light, for example, then the participant would incorrectly classify it as a non-matching image. This also led to the slightly lower accuracy results. The participants selected fewer images in general with the Large Grid (5 x 8) interface. This meant that the images that they did select were very precise, but they mis-identified a number of matching images as non-matches.

Did the Preview feature improve the use of the Swipe interface?

The Preview feature allowed the participants to see the next two images that they would be asked to categorize above the current image, although they would be slightly transparent. This gave them an opportunity to campare between multiple images if they wished. They were also able to begin the decision process for the next image before completing the current categorization action.

The addition of the Preview feature allowed the participants to complete the task slightly faster, although this difference was not significant. There were no significant differences between the two Swipe-based interfaces in terms of overall accuracy, precision, or recall. The two interfaces were almost identical in terms of overall accuracy, although the Preview Swipe interface had a higher standard error. Possible explanations for this can be seen in the participant feedback that was received.

The participants were able to provide specific comments about each of the interface styles on the TLX-style questionnaire. There were no significant differences between the two Swipe-based interfaces across any of the TLX-style measures. The participants did provide feedback comparing these two interfaces. Comments were received that stated that previewing the upcoming images was helpful when making decisions and allowed for a point of comparison. Participants reported that they liked the Preview, because it gave them more time to decide if the future images were matches, and let them move through the dataset more quickly, or get into more of a rhythm while swiping.

One drawback that the participants described with the Preview feature related to the screen size. Since there were three images displayed, as opposed to one with the regular Swipe interface, the individual images were noticeably smaller. Participants stated that this made the images harder to see, and therefore made the task more difficult. Some participants stated that they would have liked the Preview feature more on a larger screen, but that the smartphone screen was too small for it to be helpful.

Did the smaller screen change the results of the interfaces?

In Study 1, the participants were asked to complete the image classification task using a tablet. In that study the Small Grid (4 x 5) interface was clearly the most usable. It allowed the participants to complete the task the fastest, and received the best (or one of the best) ratings across every measure.

In this experiment the participants were asked to complete the task using a smartphone. This device had a much smaller screen, with less room to display the images. When using several of the interfaces (Large Grid (5 x 8), Small Grid (4 x 5), Preview Swipe) the participants commented that the pictures were too small, and it was difficult for them to identify the content of the images. This led to increased difficulty of the classification task, and, in the case of the Large Grid (5 x 8) interface, caused a few participants to complain of eye strain.

On the smaller screen, the Swipe interface was the clear winner. Participants appreciated the Preview feature in some cases, but said that even this addition was too much for the display size. Participants stated that the Preview would be a welcome addition for larger screens. The participants liked the regular Swipe interface the best because only one image was displayed at a time. This meant that the image was larger, and there weren't multiple images competing for the limited space. The RSVP interface also only displayed a single, larger image, but this interface cycled through the images too quickly. This caused the participants to feel rushed and overwhelmed.

Which interface was the best overall?

In this study, the two Swipe interfaces received the highest rankings from the participants. The participants reported that this interaction style allowed them to complete the task the fastest and the most accurately (even though these statements do not match the empirical data). They found that the Swipe interfaces were the easiest to use, and allowed them to easily categorize the specific images. They liked being able to set their own pace through the task, and this helped them to avoid feeling rushed or overwhelmed.

The interfaces all had their own benefits and drawbacks. Although the Swipe interface was the highest rated, the participants reported that the swipe interaction was sometimes uncomfortable. The Grid-based interfaces were much faster than the non-Grid interfaces, but the participants found them harder to use. The participants stated that the images in the Large Grid (5 x 8) interface were too small, and that it was hard to see the content of the images. Comments were received stating that the images were small and hard to see when using the Small Grid (4 x 5) and Preview Swipe interfaces. The Small Grid (4 x 5) and Sushi selector interfaces both performed reasonably well across the measurements, and received some positive feedback from the participants. In spite of that, neither interface measured up to the two Swipe interfaces on the small screen.

6.8.2 Comparison With Study 1

Did the task affect the performance of the interfaces?

The task that participants were asked to complete in Study 1 involved having them categorize images of leaves according to their species. This task could be difficult for people who have little to no knowledge of plant science. They may not know what features of the leaf are important in order to complete the classification process.

In order to make the study more accessible, it was decided to change the task that the participants were asked to complete. In this study, the participants were asked to categorize a more familiar image according to whether or not it matched a given prompt. The three prompts that were used were "select all images that contain a traffic light," "select all images that contain a store front," and "select all images that contain a street sign." This task was inspired by some security features on mobile websites. This research was intended to investigate whether a more familiar task would allow the participants to categorize the images more accurately.

The most clear comparison across tasks can be seen by looking at the accuracy results for the interfaces that were used in both Study 1 and Study 3(Table 6.3). These interfaces were: Small Grid (4 x 5), Swipe, Sushi, Medium Grid (5 x 8, called Large Grid in Study 3), and RSVP. It was found that the overall accuracy was slightly higher for all of these interfaces in Study 3, with the familiar task, although only the Swipe and RSVP interfaces showed a notable difference.

Table 6.3: Accuracy across tasks. Maximum of 480 possible correct responses.

Interface	Study 3	Study 1	Difference (S3-S1)
Small Grid (4 x 5)	456.22	451.90	4.32
Swipe	458.22	446.00	12.22
Sushi	455.44	450.40	5.04
Medium Grid(5 x 8)	453.11	453.05	0.06
RSVP	463.67	436.95	26.72

The precision results showed similar effects(Table 6.4). Once again, the precision was higher in Study 3 across all of the interfaces. As with the accuracy results, only the Swipe and RSVP interfaces showed a notable difference.

Table 6.4: Precision across tasks.

Interface	Study 3	Study 1	Difference (S3-S1)
Small Grid (4x5)	0.8039	0.7313	0.0726
Swipe	0.8015	0.6426	0.1589
Sushi	0.7660	0.6929	0.0731
Medium Grid(5x8)	0.8299	0.7492	0.0801
RSVP	0.7468	0.5777	0.1692

The recall results showed the opposite. Recall for every interface was lower in Study 3 than it had been in Study 1(Table 6.5). There was a difference of approximately 10% or more with every interface except for the RSVP interface.

Table 6.5: Recall across tasks.

Interface	Study 3	Study 1	Difference (S3-S1)
Small Grid (4x5)	0.7234	0.8115	-0.0881
Swipe	0.7662	0.8771	-0.1109
Sushi	0.8125	0.9188	-0.1063
Medium Grid(5x8)	0.6065	0.7698	-0.1633
RSVP	0.7879	0.8146	-0.0267

As was discussed in Chapter 4, image categorization tasks often emphasize precision over recall. This occurs in situations where incorrectly including some False images would create more of a problem than accidentally missing some True images, which is often the case when there is a large number of source images. One example of this would be while online shopping, when accidentally purchasing a product that the shopper doesn't want would clearly be more of a problem than not finding an item that would have otherwise been purchased. This isn't necessarily true when sorting objects into categories however, because the false category may be just as important as the true category.

6.9 Summary

The intention of the third study was to determine whether the results would vary if the device or task changed. Participants were asked to categorize images in a more familiar context. They were also asked to use a smartphone, since this is a device that people commonly carry with them. A controlled study with

18 participants showed that there were significant differences in the time required to complete the image categorization task with the different interfaces. Results also showed differences in accuracy, the amount of effort that the task required, and participant preference. Results showed that with smaller screen sizes, the Swipe interface is the clear winner across all of the subjective ratings.

Chapter 7

Discussion

In this chapter, all three of the experiments that were performed are discussed. It looks at the overall results concerning the impact of different interface styles on task completion time and participant effort. Interesting findings are discussed, as well as things that were learned from this research. Some of these findings led to more research questions, and possible avenues for future work. This chapter also looks at some of the limitations of the current research, and discusses how these might be resolved in future projects.

7.1 Usability

In the first study, the Small Grid (4 x 5) interface clearly outperformed several of the other styles across every measure. This interface was easy to use, allowed the participants to make decisions easily, and was not felt to require very much mental or physical effort. This interface also allowed the participants to complete the task quickly and accurately. Overall, the Small Grid (4 x 5) interface was the most highly preferred by the participants in the first study.

Different results were seen in the third study, using a smartphone. Participant feedback stated that the Grid images were small and hard to see. This led to more difficulty making categorization decisions, and caused the task to require more effort. With the smaller screen size, the Swipe interface became easiest to use. The participants still felt like the Swipe action was tedious and uncomfortable, but this was outweighed by the ability to see details in the images.

7.2 Task Completion Time

In all three of the studies, the participants were asked to complete a similar image categorization task. The first two studies asked the participants to perform fine-grained classification of leaves according to their species. The third study asked the participants to perform a more generalized task, and classify images according to whether or not they contained a common goal object, such as a street sign. All of these studies asked the participants to classify the same number of images, and the datasets contained the same ratio of matches and distractor images. The amount of time that it took for the participants to complete the task with each of the interfaces was measured in all three studies.

The first and third study explored differences between multiple interaction styles, while the second study only compared speeds of image presentation using the Sushi selector. Between the first and third study, the different interaction styles that were examined included: RSVP, Grid (in 3 sizes and with an additional Zoom feature), Sushi selector, and Swipe (with and without the added Preview). These interaction styles differ in a few key ways. The first major point of difference is whether single or multiple images were presented to the participant.

The regular Swipe and RSVP interfaces both presented the images to the participant one at a time. The other six styles showed them multiple images at once. In the first study, which used the leaf classification task, the participants stated that seeing multiple images at once was helpful, since it gave them multiple points of comparison. They were able to compare each leaf to the goal image, as well as to the other images on the screen. This sometimes helped them make their final classification decision when looking at images that they were uncertain about. This wasn't the case in the third study. Since the participants were being asked to sort images based on whether or not they contained a common and easily identifiable item (a traffic light, store front, or street sign), it wasn't as necessary to compare multiple images. The participants were able to make the classification decision based solely on the current image, and didn't need to compare across images.

In both the first and third studies, participants performed the task about twice as fast with the Grid interfaces than they did with the non-Grid interfaces. The Grid interfaces had the most images on the screen at one time. The participants were able to make categorization decisions about multiple images in parellel, which reduced the average amount of time that it took them to make a decision about each image. The participant could scan multiple images at once, and make group decisions.

Some of this effect can be seen in Study 3, when looking at the completion times for the Swipe and Preview Swipe interfaces. The task took less time to complete while using the Preview Swipe interface, although this difference was not significant. The only difference between these two interfaces was the ability to see a Preview of the next two images in line. This Preview allowed the participants to begin making the categorization decision about the future images, while reacting to the current image. This parallel processing led to the slight decrease in time.

Another difference between the interface styles is whether an action was required for every image, or just for matching images. The Swipe and Preview Swipe interfaces required the participants to react to every image. They had to swipe left for no, as well as swiping right for yes. All of the other interfaces only required the participants to react to matching images. The lack of action would cause the image to be labeled as non-matching by default. The Swipe interfaces received the highest usability scores in the third study, but was still only half the speed of the Grid interfaces. The time required for this extra action on every non-matching image adds up, and may be partly responsible for the Swipe interfaces being much slower than the Grid interfaces.

The participants reported feeling rushed with the RSVP interface, even though it took them twice as long to complete the task than it did when they used the Grid interfaces. This is another example that shows that presenting multiple images to the participant at once allows them to process or sort through multiple images simultaneously. The participants didn't report feeling rushed with any of the Grid interfaces, since they were able to set their own pace and move forward whenever they chose.

The third point of difference between the interface styles was whether or not a time limit was enforced. The RSVP interface showed each image for 1 second, and then moved to the next image automatically. The Sushi selector showed a stream of images that moved at a constant speed. The amount of time that each image was visible depended on the speed of the image stream. All of the other interfaces allowed the participant to set the pace at which they moved through the image set. The participants took about the same amount of time to complete the task with the Swipe interface and the RSVP interface in both the first and third studies. In spite of this, the TLX-style results from both studies show that the participants felt significantly more rushed while using the RSVP interface than they did while using the Swipe interface. These two interfaces were visually very similar. They both presented the participant with a single image at a time, and the image was the same size across the two interfaces. The only major difference between these interfaces is that one had a strict time limit, while the other allowed the participant to determine the speed. This shows that the enforced time limit is the cause of the participants feeling rushed. Since the overall task completion time was about the same with both interfaces, it can be seen that the participants spent longer while looking at hard to classify images. They likely moved quickly through images that obviously belonged in a certain category, and paused at images that required more examination. This explains how they could feel rushed with the RSVP interface, even though their average time per image was similar with both interfaces.

7.2.1 Image Stream Speed

After seeing some of the effects that came from enforcing a time limit in the first study, it was decided to more deeply explore one of these interface styles. The second study experimented with three different speeds of the Sushi selector interface. This was to investigate the effect that changing the speed of the image stream would have on the use of the interface. The Slow interface moved each image across the screen in 5.0 seconds, while the Medium interface took 3.9 seconds and the Fast interface took 2.7.

Participants in the first study had complained that the Sushi selector interface (which was the Slow condition in study 2) moved the images too slowly. This caused them to get bored and lose focus. In the second study the same feedback was received, which stated that the Slow interface was too slow. The participants also said that the images moved too quickly with the Fast interface. Results showed that the increased speed led to slightly lower accuracy. The participants also reported feeling more rushed and finding it harder to make categorization decisions with the Fast interface. Mixed feedback was received about the Medium interface. Some participants said that it was too fast, while others liked the speed.

When the participants were asked which interface they preferred, the results were split fairly evenly between the Slow and Medium interfaces. This showed that the ideal interface speed would be somewhere between these two interfaces. When continuing on with the third study, a speed was chosen for the Sushi selector interface that fit this ideal. The speed that was chosen showed each image for about 4.5 seconds. This speed was selected because it was approximately halfway between the Slow and Medium speeds in study 2.

7.3 Participant Perception of Speed

Studies 1 and 3 both clearly showed that all of the Grid-based interfaces allowed the participants to complete the image sort task in approximately half the time of the other interface styles. Using the Grid interfaces was objectively much faster than the other interfaces. There was also a subjective measure of speed, where participants were asked to report which interface they thought allowed them to complete the task the most quickly. In Study 3, only two of the participants reported that the Grid-based interfaces were the fastest (one each for Small and Large). Eight participants reported that the Preview Swipe interface was the fastest, and two said the Swipe interface. This means ten out of the eighteen participants (more than half) said that the Swipe-based interfaces allowed them to complete the task the most quickly. This is a very interesting result, and there are several theories that could contribute to this misconception.

Some of these theories are related to the way that human beings perceive time. Research has shown that people often perceive an apparent "speeding up" of time when in a Flow state [14] [15] [46]. Flow theory would imply that the participants in this study experienced the shortening of time while using the Swipe interface because they were experiencing a Flow state. This would mean that the Swipe interface provided the perfect balance between skill and challenge. This makes sense, because the participants were able to choose the pace of the task themselves. In order to make the task more challenging they could move faster, and they could slow down if the task became too difficult.

There is a very common saying that states that "time flies when you're having fun." Boredom can lead to an overestimation of the length of time that has passed [2]. It's possible that the participants found the Swipe interface to be more fun and engaging than the Grid interfaces. Higher levels of engagement can increase the probability of task completion, or lead to more time spent performing the task. It also makes it more likely that people will take on the task again at a later date [16]. This could imply that creating Swipe interfaces in future work could lead to a higher rate of use, and may make it easier to encourage participants to complete the required task.

It was discussed earlier that when people have more memories of an event, they perceive it to have taken more time [19]. With this task, the participants should have more memories related to using the Swipe interface than the Grid. There were more actions required, and the Swipe interface took objectively longer so they spent more time using it. This should mean that the participants would report that the Grid interface allowed them to complete the task more quickly, which is a true statement but isn't the feedback that was received.

This may be due to the amount time spent completing an action while using each interface. With the Grid interface, the participants were required to scan through and classify every image on a grid page before moving on to the next page. The only action required on a grid page was tapping once on each matching image in order to select it. The Swipe interface required an action for every single image. The participants had to decide whether or not the image matched, and then swipe left or right to confirm this choice. It has been demonstrated that participants tend to underestimate the amount of time required for actions that they caused [20] [22]. Since the participants were intentionally swiping through the image set, they may have underestimated the time that they spent reacting to each image. This slight underestimation for every image may have added up to significantly underestimating the total task time.

The difference in the number of images displayed on the screen may have influenced how the participants measured the effort or work required to complete the task. Our visual perception system is prone to errors and illusions [6] [13], which can be exploited while designing user interfaces. Humans have the tendency to view items that are displayed together as related or grouped [39]. With the Grid interface, the participants may have seen one unit of work as one grid page. They measured their speed based on the amount of time that they took to scan through and categorize all of the images on the page, before moving to the next one. With the Swipe interface one unit of work would be one single image. They measured their speed based off of the length of time that they spent categorizing each individual image. This unequal measurement of work would lead to the misconception that they moved through the images faster with the Swipe interface, when this clearly wasn't the case.

This movement could have created the illusion of faster progression through the imageset. Each time that the participants swiped through an image, the screen changed and a new image became visible. This allowed them to clearly see that they were progressing, or moving forward through the image set. With the Grid interfaces, the screen only changed after the participant finished categorizing every image on that page (20, 40, or 80). The participants took longer to feel like they were making any progress. This may have led to the participants feeling like the task moved more slowly when using the Grid interfaces.

This illusion of progress can be related to the concept of psychological myopia. This refers to the tendency to be short-sighted while making a decision, or to ignore background information while focusing only on information directly related to their current decision. One example of psychological myopia is known as the advantage effect, or the illusion of advantage, and is described in Hsee *et al.'s* research into Medium Maximization. In this research a "medium" refers to some type of tokens or points which can be exchanged in order to receive a desirable reward or outcome [25]. The advantage effect described here concerns situations in which people have a choice between an action that requires more effort and an action that requires less effort, when the outcomes of each action are similar or neither has a large advantage. The research shows that if a "medium" - which in this research referred to tokens or points with no inherent value, but that can be exchanged for prizes or desirable outcomes - is provided in much higher quantities in response to the high-effort task than the low-effort task, participants will see this as an advantage and choose the high-effort task over the easier one. In this study, the amount of perceived progress towards the goal acted as a medium. The participants preferred the higher effort action of swiping (which was described as tedious and uncomfortable), because it allowed them to feel as if they were making more progress, and moving forward more quickly.

This also relates to a phenomenon known as the endowed progress effect [33]. This effect shows that when people are given artificial advancement towards a goal, they show more persistence in achieving that goal. For example, someone who receives a punch card that allows them to earn a free coffee after buying 10 which has 2 punches already, and only 8 more remaining to the goal, will be more persistent than someone who receives an empty card allowing them to earn a free coffee after 8 purchases. Even though both cards require 8 more purchases to receive the reward, the person in the first case will fill the card more quickly and make more frequent purchases [33].

Receiving artificial progress also makes someone more likely to continue towards the goal, or less likely to abandon the required task [33]. This effect implies that since the actions involved in the Swipe interface provide an illusion that more progress is being made with each action, people using this interface would be more likely to complete a task. They would also be more motivated, and would be willing to devote more time to the task. This is an avenue that could be explored in future work, as it may have impacts on user retention and motivation.

7.4 Which Interface is the Best?

It was found that the Small Grid (4 x 5) interface was by far the best based on the criteria that was measured in the first study. The participants completed the task much more quickly with all of the Grid-based interfaces. The Small Grid (4 x 5) interface received the best or one of the best scores across every measure in the TLX-style results for the first study. The participants found it to be easy to use, and it was the most highly preferred interface. This could be because many participants are already very familiar with navigating grids of images or icons. This is required in several every day tasks such as using an image search engine or shopping online, or even just finding and opening a specific app on a smartphone's homescreen.

In the third study, the Swipe-based interfaces (Swipe and Preview Swipe) received the highest preference rankings. The Swipe interface received the best score across almost every TLX-style measure. This is in spite of the fact that the participants reported that the Swipe action was more tiring, and was tedious to perform repetitively.

The two major differences between these studies were the task that the participants were asked to perform, and the device that they were required to use. Changing the task from the leaf categorization task used in the first study to the CAPTCHA task used in the third study doesn't really explain why a different interface type was preferred. The most likely explanation for this change comes from the smaller screen size in study

3.

The first two studies were performed on a Microsoft Surface Pro 4 tablet, with a 12.3-inch display. The third study used a Nexus 5 smartphone with a 4.95-inch display. In the first study, feedback was received stating that the images were too small when using the Large Grid (8 x 10) interface, and that this made the task more difficult. In Study 3, the participants stated that the images were small and hard to see when using both of the Grid interfaces.

Overall, this research confirmed that grids are the best interface as far as speed is concerned. The screen size of the device can have an effect on which interface would be the best choice for a task. The image size needs to be taken into account if images are complex or detailed, as small images can make the task overwhelming or difficult. The Swipe action was described as uncomfortable, but on a phone this interface still seems to be the best choice for making detailed categorization decisions.

It is also possible that the best interface for some tasks may be a combination of the interfaces that were created. For example, smart phone homescreens often use a combination of Grid and Swipe interfaces to display the applications available on the phone. The user swipes back and forth to navigate through pages of apps in a grid layout. Some phones even allow the user to create folders of apps, which is displayed as a smaller grid laid out in the main grid square. The image size is not a problem in this context, because many of the icons are very distinctive. They often use bright colours and simple shapes, and seldom contain many small details.

Chapter 8

Conclusions

Image categorization tasks are extremely common, and can take many different forms. Some examples of these tasks can include sorting phenotypic images into groups, or searching through profiles on a dating app. There are many possible interface designs that can be used to carry out these tasks, and little previous research on how the selection technique can impact the user's experience and performance.

In this thesis we described three studies that were performed in order to investigate the impact of the interface style on the user experience when completing an image classification task. Some important factors and differences between the different selection styles were identified. Implications of some of these factors were discussed, and it was explained how they impacted the usability of the interface. The Small Grid (4 x 5) and Swipe interfaces were identified as the clear winners in these two studies. Benefits and drawbacks of both of these styles were presented, as well as types of situations in which they would each be ideal.

The contributions of this work include identifying important design factors to take into consideration when creating interfaces for an image classification task. Empirical evidence of the impact and effects that these factors have on usability, performance, and user preference was provided. Lastly, the Small Grid (4 x 5) and Swipe interfaces were identified as the clear winners in the different situations, and strengths and weaknesses of each of these interfaces were described. The results suggested that grid interfaces allow for fast and easy image classification, but that when the images contain a lot of detail or the the screen is small the grid images can be difficult to see. Swipe interfaces often display a single, large image, which allows the user to see details even when screen space is limited.

In the third study, several research questions were answered that had been raised by earlier experiments. Participants were asked to complete a more familiar image categorization task, in order to see if this would improve accuracy. They were also asked to use a smartphone instead of a tablet, since this is a more common and portable device and this allowed for investigation into the effects of a smaller screen. A controlled study showed that there were significant differences in the time and effort required to complete the task with different interfaces. This research also showed differences in accuracy and user preference. Results showed that the Swipe interface was the most suited to the smaller screen size.

The second study investigated the impact of changing the image-stream speed in the Sushi selector interface. Participant performance and preference was examined across three different image-stream speeds. It was found that the Fast speed was the worst on all counts. It was discovered that the ideal speed would be somewhere between the Slow and Medium speeds used in this study. Moving forward a speed that would display images for approximately 4.5 seconds was used, since this was about halfway between these two speeds.

The initial investigation into the usability of different image sort interfaces was performed on a tablet. Participants completed a leaf sorting task using seven different interfaces. Significant differences were found in user preference, subjective effort, task completion time and accuracy across the different interfaces. The Small Grid interface outperformed the other styles in every category, and was preferred by the participants. Based on the criteria that was measured, the Small Grid (4 x 5) interface was the clear winner. Therefore, we recommend that app developers adopt small-grid style interfaces for image categorization tasks when space allows and consider swipe style interfaces on small screens, where the reduced size of each individual image makes grids impractical.

8.1 Limitations and Future Work

This research led to several interesting findings, and raised more research questions that would be interesting to investigate. The initial results showed that Grid and Swipe based interfaces both performed very well in certain situations, and both have their own benefits and drawbacks. We would like to continue investigating the use of these interfaces in new contexts, and possibly look at the effect of combining some of the features of these two styles in order to create the best possible interface.

One result showed that many participants felt like they were able to complete the task more quickly while using the Swipe interface, even though the results showed that this was an illusion. Several possible theories were discussed that may have contributed to this effect. It would be interesting to look more closely at these theories, and see if the participants were more engaged while using the Swipe interface. It would also be interesting to investigate whether this interface style may cause participants to decide to spend more time classifying images, or choose to return to the task more often.

One limitation with this research was the need to be able to measure the accuracy of the participant responses. This meant that the studies were limited to categorization tasks that had a “right” or “wrong” answer for every image. Many common image categorization tasks, such as online dating, choosing images for a presentation, or deciding on which pet to adopt, have subjective categories. The criteria for a match is determined by the user, and their feelings about their decisions are the important measure of accuracy.

Future work would investigate which interfaces are most preferred and the easiest to use when completing a subjective image categorization task, with no real wrong answers. It could also look into how satisfied the participants are with their choices after completing the task, and whether or not this is influenced by the interface style. This line of investigation also ties into the “paradox of choice,” which suggests that having fewer choices can actually reduce anxiety and lead to more satisfaction with the final decision. We are interested in discovering whether changing the interface used to make the choice will lead to this effect.

It would also be interesting to look at tasks that don't involve binary classification. This could include sorting items into several categories, or into a ranked list. There are some limitations introduced by adding additional categories. The gestures for each category need to be distinct enough to prevent frequent errors. An example would be a Swipe interface with different categories assigned to Up, Down, Left, and Right swipe directions. Some interface styles (RSVP, for example) don't allow for multiple unique gestures and would be ill suited to tasks that require sorting into multiple categories.

This research was motivated in part by the need for systems that can automatically classify images in several other research projects that our colleagues are working on. The machine learning algorithms in these projects will require huge, labeled training sets in order to begin accurately classifying the test images. These training sets will need to be labeled initially by humans, in order to even begin the process of automatic classification. Some of the test datasets may also require participants to narrow down or confirm the labels that are assigned by the classifier, as machine learning algorithms are often less accurate than human-assigned labels. Moving forward this system could be adapted so that it can be used to assign labels to these images.

This will include integrating the imagesets required for these other projects. One way to quickly label a large number of images would be to use crowdsourcing methods, in order to allow more people to take part in the task. This may allow for labeling the datasets more quickly and with less cost. Research could further investigate the accuracy of crowdsourced labels for the image sort tasks that will need to be performed.

Another possible avenue for future work would be to test this system with a wider variety of participants. All of the participants in these studies were found through the University website, and were students at the University of Saskatchewan. University students do not provide an accurate representation of the wider population. Testing the interfaces with more participants would allow for a more diverse sample set, and a more accurate representation of the general population. This would also allow for analysis of the results based on different demographic groups. We might see that factors like age or occupation influence interface preference or performance results.

References

[1] Gaurav Agarwal, Peter Belhumeur, and Steven Feiner et al. First steps toward an electronic field guide for plants. *Taxon*, 55(3):597–610, August 2006. `http://www.jstor.org/stable/25065637`, last accessed on 06/25/20.

[2] Ava-Ann A. Allman and James A. Danckert. Time flies when you're having fun: Temporal estimation and the experience of boredom. *Brain and Cognition*, 59:236–245, 2005.

[3] Rosemarie Anderson, Sam T. Manoogian, and J. Steven Reznick. The undermining and enhancing of intrinsic motivation in preschool children. *Journal of Personality and Social Psychology*, 34(5):915–922, 1976.

[4] Peter N. Belhumeur, Daozheng Chen, and Steven Feiner et al. Searching the world's herbaria: A system for visual identification of plant species. *Lecture Notes in Computer Science Computer Vision - ECCV*, pages 116–129, 2008. `http://doi.org/10.1007/978-3-540-88693-8_9`, last accessed on 06/25/20.

[5] Nima Bigdely-Shamlo, Andrey Vankov, Rey R. Ramirez, and Scott Makeig. Brain activity-based image classification from rapid serial visual presentation. *IEEE Trans Neural Syst. Rehabil. Eng. IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 16(5):432–441, 2008. `http://doi.org/10.1109/tnsre.2008.2003381`, last accessed on 06/25/20.

[6] J Richard Block and Harold Yuker. *Can you believe your eyes?* Routledge, 2013.

[7] Steve Branson, Catherine Wah, and Florian Schroff et al. Visual recognition with humans in the loop. *Computer Vision - ECCVLecture Notes in Computer Science*, pages 438 – 451, 2010. `http://doi.org/10.1007/978-3-642-15561-1_32`, last accessed on 06/25/20.

[8] Donald E. Broadbent and Margaret H. P. Broadbent. From detection to identification: Response to multiple targets in rapid serial visual presentation. *Perception & Psychophysics*, 42(2):105–113, 1987. `http://doi.org/10.3758/bf03210498`, last accessed on 06/25/20.

[9] Oscar De Bruijn and Robert Spence. Rapid serial visual presentation. In *Proceedings of the working conference on Advanced visual interfaces - AVI '00*, 2000.

[10] Monica Chew and J. D. Tygar. Image recognition captchas. *Lecture Notes in Computer Science Information Security:*, pages 268 – 279, 2004. `http://doi.org/10.1007/978-3-540-30144-8_23`, last accessed on 06/25/20.

[11] Sung-Jung Cho, Roderick Murray-Smith, and Yeun-Bae Kim. Multi-context photo browsing on mobile devices based on tilt dynamics. *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services - MobileHCI '07*, 2007. `http://doi.org/10.1145/1377999.1378006`, last accessed on 06/25/20.

[12] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popovic, and Foldit players. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 08 2010. `http://dx.doi.org/10.1038/nature09304`, last accessed on 06/25/20.

[13] Stanley Coren and Joan Stern Girgus. *Seeing is deceiving: The psychology of visual illusions.* JSTOR, 1978.

[14] Mihaly Csikszentmihalyi. *Beyond boredom and anxiety*. San Francisco: Jossey-Bass, 2000(Original work published 1975).

[15] Mihaly Csikszentmihalyi and Jeanne Nakamura. The concept of flow. *Flow ans the Foundations of Positive Psychology*, pages 239–263, 2014.

[16] E. Deci and R. Ryan. *Intrinsic motivation and self-determination in human behavior*. New York: Plenum, 1985.

[17] John R. Douceur, Jeremy Elson, Jon Howell, and Jared Saul. Asirra: A captcha that exploits interest-aligned manual image categorization. *ACM*, 2007.

[18] Sylvie Droit-Volet and Warren H. Meck. How emotions colour our perception of time. *TRENDS in Cognitive Sciences*, 11(12):504–513, 2007.

[19] David M. Eagleman. Human time perception and its illusions. *Current Opinion in Neurobiology*, 18:131–136, 2008.

[20] David M. Eagleman and Alex O. Holcombe. Causality and the perception of time. *TRENDS in Cognitive Sciences*, 6(8):323–325, 2002.

[21] Freepik.com. Freepik. `https://www.freepik.com/`, image: Freepik.com, last accessed on 07/01/20.

[22] Patrick Haggard and Jonathan Cole. Intention, attention and the temporal experience of action. *Consciousness and Cognition*, 16:211–220, 2007.

[23] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in Psychology Human Mental Workload:*, pages 139–183, 1988. `http://doi.org/10.1016/s0166-4115(08)62386-9`, last accessed on 06/25/20.

[24] Uta Hinrichs, Sheelagh Carpendale, Stacey D. Scott, and Eric Pattison. Interface currents: Supporting fluent collaboration on tabletop displays. *Smart Graphics Lecture Notes in Computer Science*, pages 185–197, 2005. `http://doi.org/10.1007/11536482_16`, last accessed on 06/25/20.

[25] Christopher K. Hsee, Fang Yu, Jiao Zhang, and Yan Zhang. Medium maximization. *Journal of Consumer Research*, 30(1):1–14, 2003. `http://www.jstor.org/stable/10.1086/374702`.

[26] Google Inc. Google web search. `https://www.google.com`, last accessed on 06/25/20.

[27] Richard B. Ivry and John E. Schlerf. Dedicated and intrinsic models of time perception. *Trends in Cognitive Sciences*, 12(7):273–280, 2008.

[28] Marshall G. Jones. Creating electronic learning environments: Games, flow, and the user interface. *In:Procedings of Selected Research and Development Presentations at the National Convention of the Asociation for Educational Comunications and Technology(AECT)*, pages 204–214, 1998.

[29] Neeraj Kumar, Peter N. Belhumeur, and Arijit Biswas et al. Leafsnap: A computer vision system for automatic plant species identification. *Computer Vision - ECCVLecture Notes in Computer Science*, pages 502–516, 2012. `http://doi.org/10.1007/978-3-642-33709-3_36`, last accessed on 06/25/20.

[30] Warren H. Meck. Neuropsychology of timing and time perception. *Brain and Cognition*, 58:1–8, 2005.

[31] Raju Mudhar. Tinder's swipe interface gets swiped by other apps. *thestar.com*, 2014. `https://www.thestar.com/life/technology/2014/08/06/tinders_swipe_interface_gets_swiped_by_other_apps.html`, last accessed on 06/25/20.

[32] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 2008. `http://doi.org/10.1109/icvgip.2008.47`, last accessed on 06/25/20.

[33] Joseph C. Nunes and Xavier Drèze. The endowed progress effect: How artificial advancement increases effort. *Journal of Consumer Research*, 32:504–512, 2006.

[34] Massimiliano Oliveri, Carmelo Mario Vicario, Silvia Salerno, Giacomo Koch, Patrizia Turriziani, Renata Mangano, Gaetana Chillemi, and Carlo Caltagirone. Perceiving numbers alters time perception. *Neuroscience Letters*, 438:308–311, 2008.

[35] Jared Saul. Petfinder. `https://www.petfinder.com`, last accessed on 06/25/20.

[36] Neil Savage. Gaining wisdom from crowds. *Communications of the ACM*, 55(3):13–15, 2012.

[37] T3. Hands-free tinder. 2015. `https://youtu.be/bd0kuwi9cle`, last accessed on 06/25/20.

[38] Tinder. Tinder. `https://www.gotinder.com`, last accessed on 06/25/20.

[39] D. Todorovic. Gestalt principles. *Scholarpedia*, 3(12):5345, 2008. revision #91314.

[40] Lous von Ahn, Manuel Blum, Nicholas Hopper, and John Langford. Captcha. `https://www.captcha.net`, last accessed on 06/25/20.

[41] Luis von Ahn. Massive-scale online collaboration. *TEDxCMU*. `http://www.ted.com/talks/luis_von_ahn_massive_scale_online_collaboration.html`, last accessed on 06/25/20.

[42] Luis von Ahn, Manuel Blum, and John Langford. Telling humans and computers apart automatically. *Comminications of the ACM*, 47(2):57–60, 2004.

[43] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the 2004 conference on Human factors in computing systems - CHI '04*, 2004.

[44] Luis von Ahn, Ruoran Liu, and Manuel Blum. Peekaboom. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, 2006.

[45] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. recaptcha: Human-based character recognition via web security measures. *Science*, 321:1465–1468, 2008. `www.sciencemag.org`.

[46] Jane Webster, Linda Klebe Trevino, and Lisa Ryan. The dimensionality and correlates of flow in human-computer interactions. *Computers in Human Behavior*, 9:411–426, 1993.

Appendix A

Demographics Form

Participant ID:

Sex:
Male
Female
Other, please specify...
Prefer not to say

Are you a student?
Yes No
If yes, what is your field of study?

As far as you are aware, do you have any color vision deficiencies (e.g., color blindness)?
Yes No
If yes, please specify:

On average, how much time do you spend on computers per day?
Less than 30 minutes
30 - 60 minutes
1 - 2 hours
2 - 4 hours
4 - 8 hours
More than 8 hours

How much time do you spend playing computer, video, or console games?
None
Less than 3 hours a week
3 - 7 hours a week
1 - 2 hours a day
More than 2 hours a day

How often do you use touch screen devices (i.e. smartphones, tablets, etc.)?
Never
Less than 3 hours a week
3 - 7 hours a week
1 - 2 hours a day
More than 2 hours a day

How often do you play games on touch screen devices?
Never
Less than 3 hours a week
3 - 7 hours a week
1 - 2 hours a day
More than 2 hours a day

Please list the games you play on touch screen devices:

How often do you use a tablet?

Never
Less than 3 hours a week
3 - 7 hours a week
1 - 2 hours a day
More than 2 hours a day

How much time do you spend using a word processor, email, or instant messaging?
None
Less than 3 hours a week
3 - 7 hours a week
1 - 2 hours a day
More than 2 hours a day

Appendix B

TLX-Style Questionnaire

Participant ID:

Interface Style:

This style was easy to use:
Strongly disagree
Disagree somewhat
Neither agree nor disagree
Agree somewhat
Strongly agree

What aspects of this style made it easy or hard to use?

It was easy to make decisions about a specific image with this style:
Strongly disagree
Disagree somewhat
Neither agree nor disagree
Agree somewhat
Strongly agree

I felt overwhelmed with the number of choices in this style:
Strongly disagree
Disagree somewhat
Neither agree nor disagree
Agree somewhat
Strongly agree

I felt felt rushed while making decisions with this style:
Strongly disagree
Disagree somewhat
Neither agree nor disagree
Agree somewhat
Strongly agree

I was able to accurately identify matching images using this style:
Strongly disagree
Disagree somewhat
Neither agree nor disagree
Agree somewhat
Strongly agree

The interface was visually appealing:
Strongly disagree
Disagree somewhat
Neither agree nor disagree
Agree somewhat
Strongly agree

Why or why not?

Using this selection style was mentally demanding:
Strongly disagree
Disagree somewhat
Neither agree nor disagree
Agree somewhat
Strongly agree
Why or why not?

Using this selection style was physically demanding:
Strongly disagree
Disagree somewhat
Neither agree nor disagree
Agree somewhat
Strongly agree

Why or why not?

Any additional comments about this style?

Appendix C

Summary Questionnaires

C.1 Study 1

Participant ID:

Please rank the seven selection styles in order of preference (1 = best, 2 = worst):
Small Grid
Swipe
Medium Grid
Sushi
Large Grid
Zoom Grid
RSVP

With which selection style were you the fastest?:

With which selection style did you make the fewest errors?:

C.2 Study 2

Participant ID:

Please rank the three selection styles in order of preference (1 = best, 2 = worst):
Slow
Medium
Fast

With which interface did you make the fewest errors?:

Do you have any additional comments about any of the interfaces?

C.3 Study 3

Participant ID:

Please rank the six selection styles in order of preference (1 = best, 6 = worst):
Small Grid
Swipe
Sushi
Large Grid
Preview Swipe
RSVP

With which selection style were you the fastest?:

With which selection style did you make the fewest errors?:

Appendix D

ANOVA Results - Study 1

Table D.1: TLX ANOVA results for Study 1. Significant results are starred.

		Sum of Squares	df	Mean Square	F	Sig.
Easy To Use	Between Groups	59.400	6	9.900	8.128	0.00*
	Within Groups	162.000	133	1.218		
	Total	221.400	139			
Easy Decisions	Between Groups	44.571	6	7.429	6.416	0.00*
	Within Groups	154.000	133	1.158		
	Total	198.571	139			
Overwhelmed	Between Groups	13.086	6	2.181	1.127	.350
	Within Groups	257.450	133	1.936		
	Total	270.536	139			
Rushed	Between Groups	86.271	6	14.379	10.453	0.00*
	Within Groups	182.950	133	1.376		
	Total	269.221	139			
Subjective Accuracy	Between Groups	18.271	6	3.045	2.854	.012*
	Within Groups	141.900	133	1.067		
	Total	160.171	139			
Visual Appeal	Between Groups	17.600	6	2.933	2.653	.018*
	Within Groups	147.050	133	1.106		
	Total	164.650	139			
Mental Demand	Between Groups	39.600	6	6.600	4.012	0.001*
	Within Groups	218.800	133	1.645		
	Total	258.400	139			
Physical Demand	Between Groups	10.371	6	1.729	1.246	0.287
	Within Groups	184.450	133	1.387		
	Total	194.821	139			

Appendix E

Tukey Results - Study 1

Table E.1: Time Tukey HSD results for Study 1. Significant results are starred.

Dependent Variable	(I) Interface Style	(J) Interface Style	Mean Difference (I-J)	Sig.
Time	Sm. Grid (4x5)	Swipe	-37.26406758000*	0
		Md. Grid (5x8)	8.72499515	0.28
		Sushi	-39.85657283000*	0
		Lg. Grid (8x10)	13.20714758000*	0.015
		Zoom Grid (8x10)	4.069429585	0.942
		RSVP	-37.55523858000*	0
	Swipe	Sm. Grid (4x5)	37.26406758000*	0
		Md. Grid (5x8)	45.98906273000*	0
		Sushi	-2.592505252	0.994
		Lg. Grid (8x10)	50.47121516000*	0
		Zoom Grid (8x10)	41.33349716000*	0
		RSVP	-0.291171002	1
	Md. Grid (5x8)	Sm. Grid (4x5)	-8.72499515	0.28
		Swipe	-45.98906273000*	0
		Sushi	-48.58156798000*	0
		Lg. Grid (8x10)	4.482152425	0.91
		Zoom Grid (8x10)	-4.655565565	0.894
		RSVP	-46.28023373000*	0
	Sushi	Sm. Grid (4x5)	39.85657283000*	0
		Swipe	2.592505252	0.994
		Md. Grid (5x8)	48.58156798000*	0
		Lg. Grid (8x10)	53.06372041000*	0
		Zoom Grid (8x10)	43.92600242000*	0
		RSVP	2.30133425	0.997
	Lg. Grid (8x10)	Sm. Grid (4x5)	-13.20714758000*	0.015
		Swipe	-50.47121516000*	0
		Md. Grid (5x8)	-4.482152425	0.91
		Sushi	-53.06372041000*	0
		Zoom Grid (8x10)	-9.137717991	0.229
		RSVP	-50.76238616000*	0
	Zoom Grid (8x10)	Sm. Grid (4x5)	-4.069429585	0.942
		Swipe	-41.33349716000*	0
		Md. Grid (5x8)	4.655565565	0.894
		Sushi	-43.92600242000*	0
		Lg. Grid (8x10)	9.137717991	0.229
		RSVP	-41.62466817000*	0
	RSVP	Sm. Grid (4x5)	37.55523858000*	0
		Swipe	0.291171002	1
		Md. Grid (5x8)	46.28023373000*	0
		Sushi	-2.30133425	0.997
		Lg. Grid (8x10)	50.76238616000*	0
		Zoom Grid (8x10)	41.62466817000*	0

Table E.2: Accuracy Tukey HSD results for Study 1. Significant results are starred.

Dependent Variable	(I) Interface Style	(J) Interface Style	Mean Difference (I-J)	Sig.
Correct Responses	Sm.	Swipe	5.9	0.971
		Md. Grid (5x8)	-1.15	1
		Sushi	1.5	1
		Lg. Grid (8x10)	5	0.988
		Zoom Grid (8x10)	-1.3	1
		RSVP	14.95	0.256
	Swipe	Sm. Grid (4x5)	-5.9	0.971
		Md. Grid (5x8)	-7.05	0.933
		Sushi	-4.4	0.994
		Lg. Grid (8x10)	-0.9	1
		Zoom Grid (8x10)	-7.2	0.926
		RSVP	9.05	0.808
	Md. Grid (5x8)	Sm. Grid (4x5)	1.15	1
		Swipe	7.05	0.933
		Sushi	2.65	1
		Lg. Grid (8x10)	6.15	0.965
		Zoom Grid (8x10)	-0.15	1
		RSVP	16.1	0.18
	Sushi	Sm. Grid (4x5)	-1.5	1
		Swipe	4.4	0.994
		Md. Grid (5x8)	-2.65	1
		Lg. Grid (8x10)	3.5	0.998
		Zoom Grid (8x10)	-2.8	1
		RSVP	13.45	0.382
	Lg. Grid (8x10)	Sm. Grid (4x5)	-5	0.988
		Swipe	0.9	1
		Md. Grid (5x8)	-6.15	0.965
		Sushi	-3.5	0.998
		Zoom Grid (8x10)	-6.3	0.961
		RSVP	9.95	0.73
	Zoom Grid (8x10)	Sm. Grid (4x5)	1.3	1
		Swipe	7.2	0.926
		Md. Grid (5x8)	0.15	1
		Sushi	2.8	1
		Lg. Grid (8x10)	6.3	0.961
		RSVP	16.25	0.171
	RSVP	Sm. Grid (4x5)	-14.95	0.256
		Swipe	-9.05	0.808
		Md. Grid (5x8)	-16.1	0.18
		Sushi	-13.45	0.382
		Lg. Grid (8x10)	-9.95	0.73
		Zoom Grid (8x10)	-16.25	0.171
Precision	Sm. Grid (4x5)	Swipe	0.088688131	0.714
		Md. Grid (5x8)	-0.017895305	1
		Sushi	0.038387487	0.994
		Lg. Grid (8x10)	0.005240118	1
		Zoom Grid (8x10)	-0.044949861	0.986
		RSVP	0.153636347	0.11
	Swipe	Sm. Grid (4x5)	-0.088688131	0.714
		Md. Grid (5x8)	-0.106583435	0.508
		Sushi	-0.050300644	0.975

		Lg. Grid (8x10)	-0.083448013	0.768
		Zoom Grid (8x10)	-0.133637992	0.235
		RSVP	0.064948217	0.916
	Md. Grid (5x8)	Sm. Grid (4x5)	0.017895305	1
		Swipe	0.106583435	0.508
		Sushi	0.056282791	0.957
		Lg. Grid (8x10)	0.023135422	1
		Zoom Grid (8x10)	-0.027054556	0.999
		RSVP	.171531652000*	0.049
	Sushi	Sm. Grid (4x5)	-0.038387487	0.994
		Swipe	0.050300644	0.975
		Md. Grid (5x8)	-0.056282791	0.957
		Lg. Grid (8x10)	-0.033147369	0.997
		Zoom Grid (8x10)	-0.083337348	0.769
		RSVP	0.115248861	0.41
	Lg. Grid (8x10)	Sm. Grid (4x5)	-0.005240118	1
		Swipe	0.083448013	0.768
		Md. Grid (5x8)	-0.023135422	1
		Sushi	0.033147369	0.997
		Zoom Grid (8x10)	-0.050189979	0.975
		RSVP	0.14839623	0.136
	Zoom Grid (8x10)	Sm. Grid (4x5)	0.044949861	0.986
		Swipe	0.133637992	0.235
		Md. Grid (5x8)	0.027054556	0.999
		Sushi	0.083337348	0.769
		Lg. Grid (8x10)	0.050189979	0.975
		RSVP	.198586208000*	0.012
	RSVP	Sm. Grid (4x5)	-0.153636347	0.11
		Swipe	-0.064948217	0.916
		Md. Grid (5x8)	-.171531652000*	0.049
		Sushi	-0.115248861	0.41
		Lg. Grid (8x10)	-0.14839623	0.136
		Zoom Grid (8x10)	-.198586208000*	0.012
Recall	Sm. Grid (4x5)	Swipe	-0.065625	0.819
		Md. Grid (5x8)	0.041666667	0.977
		Sushi	-0.107291667	0.286
		Lg. Grid (8x10)	0.107291667	0.286
		Zoom Grid (8x10)	0.078125	0.667
		RSVP	-0.003125	1
	Swipe	Sm. Grid (4x5)	0.065625	0.819
		Md. Grid (5x8)	0.107291667	0.286
		Sushi	-0.041666667	0.977
		Lg. Grid (8x10)	.172916667000*	0.008
		Zoom Grid (8x10)	0.14375	0.051
		RSVP	0.0625	0.851
	Md. Grid (5x8)	Sm. Grid (4x5)	-0.041666667	0.977
		Swipe	-0.107291667	0.286
		Sushi	-.148958333000*	0.037
		Lg. Grid (8x10)	0.065625	0.819
		Zoom Grid (8x10)	0.036458333	0.988
		RSVP	-0.044791667	0.967
	Sushi	Sm. Grid (4x5)	0.107291667	0.286
		Swipe	0.041666667	0.977
		Md. Grid (5x8)	.148958333000*	0.037

		Lg. Grid (8x10)	.214583333000*	0
		Zoom Grid (8x10)	.185416667000*	0.003
		RSVP	0.104166667	0.321
	Lg. Grid (8x10)	Sm. Grid (4x5)	-0.107291667	0.286
		Swipe	-.172916667000*	0.008
		Md. Grid (5x8)	-0.065625	0.819
		Sushi	-.214583333000*	0
		Zoom Grid (8x10)	-0.029166667	0.997
		RSVP	-0.110416667	0.254
	Zoom Grid (8x10)	Sm. Grid (4x5)	-0.078125	0.667
		Swipe	-0.14375	0.051
		Md. Grid (5x8)	-0.036458333	0.988
		Sushi	-.185416667000*	0.003
		Lg. Grid (8x10)	0.029166667	0.997
		RSVP	-0.08125	0.624
	RSVP	Sm. Grid (4x5)	0.003125	1
		Swipe	-0.0625	0.851
		Md. Grid (5x8)	0.044791667	0.967
		Sushi	-0.104166667	0.321
		Lg. Grid (8x10)	0.110416667	0.254
		Zoom Grid (8x10)	0.08125	0.624

Table E.3: TLX Tukey HSD results for Study 1. Significant results are starred.

Dependent Variable	(I) Interface Style	(J) Interface Style	Mean Difference (I-J)	Sig.
Easy to Use	Sm. Grid (4x5)	Swipe	0.05	1
		Md. Grid (5x8)	0.35	0.953
		Sushi	0.4	0.912
		Lg. Grid (8x10)	1.250*	0.008
		Zoom Grid (8x10)	0	1
		RSVP	1.800*	0
	Swipe	Sm. Grid (4x5)	-0.05	1
		Md. Grid (5x8)	0.3	0.978
		Sushi	0.35	0.953
		Lg. Grid (8x10)	1.200*	0.013
		Zoom Grid (8x10)	-0.05	1
		RSVP	1.750*	0
	Md. Grid (5x8)	Sm. Grid (4x5)	-0.35	0.953
		Swipe	-0.3	0.978
		Sushi	0.05	1
		Lg. Grid (8x10)	0.9	0.141
		Zoom Grid (8x10)	-0.35	0.953
		RSVP	1.450*	0.001
	Sushi	Sm. Grid (4x5)	-0.4	0.912
		Swipe	-0.35	0.953
		Md. Grid (5x8)	-0.05	1
		Lg. Grid (8x10)	0.85	0.192
		Zoom Grid (8x10)	-0.4	0.912
		RSVP	1.400*	0.002
	Lg. Grid (8x10)	Sm. Grid (4x5)	-1.250*	0.008
		Swipe	-1.200*	0.013
		Md. Grid (5x8)	-0.9	0.141
		Sushi	-0.85	0.192

		Zoom Grid (8x10)	-1.250*	0.008
		RSVP	0.55	0.698
	Zoom Grid (8x10)	Sm. Grid (4x5)	0	1
		Swipe	0.05	1
		Md. Grid (5x8)	0.35	0.953
		Sushi	0.4	0.912
		Lg. Grid (8x10)	1.250*	0.008
		RSVP	1.800*	0
	RSVP	Sm. Grid (4x5)	-1.800*	0
		Swipe	-1.750*	0
		Md. Grid (5x8)	-1.450*	0.001
		Sushi	-1.400*	0.002
		Lg. Grid (8x10)	-0.55	0.698
		Zoom Grid (8x10)	-1.800*	0
EasyDecisions	Sm. Grid (4x5)	Swipe	0.25	0.99
		Md. Grid (5x8)	0.45	0.84
		Sushi	0.3	0.975
		Lg. Grid (8x10)	1.150*	0.016
		Zoom Grid (8x10)	0	1
		RSVP	1.600*	0
	Swipe	Sm. Grid (4x5)	-0.25	0.99
		Md. Grid (5x8)	0.2	0.997
		Sushi	0.05	1
		Lg. Grid (8x10)	0.9	0.121
		Zoom Grid (8x10)	-0.25	0.99
		RSVP	1.350*	0.002
	Md. Grid (5x8)	Sm. Grid (4x5)	-0.45	0.84
		Swipe	-0.2	0.997
		Sushi	-0.15	0.999
		Lg. Grid (8x10)	0.7	0.384
		Zoom Grid (8x10)	-0.45	0.84
		RSVP	1.150*	0.016
	Sushi	Sm. Grid (4x5)	-0.3	0.975
		Swipe	-0.05	1
		Md. Grid (5x8)	0.15	0.999
		Lg. Grid (8x10)	0.85	0.168
		Zoom Grid (8x10)	-0.3	0.975
		RSVP	1.300*	0.004
	Lg. Grid (8x10)	Sm. Grid (4x5)	-1.150*	0.016
		Swipe	-0.9	0.121
		Md. Grid (5x8)	-0.7	0.384
		Sushi	-0.85	0.168
		Zoom Grid (8x10)	-1.150*	0.016
		RSVP	0.45	0.84
	Zoom Grid (8x10)	Sm. Grid (4x5)	0	1
		Swipe	0.25	0.99
		Md. Grid (5x8)	0.45	0.84
		Sushi	0.3	0.975
		Lg. Grid (8x10)	1.150*	0.016
		RSVP	1.600*	0
	RSVP	Sm. Grid (4x5)	-1.600*	0
		Swipe	-1.350*	0.002
		Md. Grid (5x8)	-1.150*	0.016
		Sushi	-1.300*	0.004

		Lg. Grid (8x10)	-0.45	0.84
		Zoom Grid (8x10)	-1.600*	0
Overwhelmed	Sm. Grid (4x5)	Swipe	-0.25	0.998
		Md. Grid (5x8)	-0.4	0.971
		Sushi	-0.3	0.993
		Lg. Grid (8x10)	-1.05	0.213
		Zoom Grid (8x10)	-0.6	0.82
		RSVP	-0.35	0.985
	Swipe	Sm. Grid (4x5)	0.25	0.998
		Md. Grid (5x8)	-0.15	1
		Sushi	-0.05	1
		Lg. Grid (8x10)	-0.8	0.538
		Zoom Grid (8x10)	-0.35	0.985
		RSVP	-0.1	1
	Md. Grid (5x8)	Sm. Grid (4x5)	0.4	0.971
		Swipe	0.15	1
		Sushi	0.1	1
		Lg. Grid (8x10)	-0.65	0.758
		Zoom Grid (8x10)	-0.2	0.999
		RSVP	0.05	1
	Sushi	Sm. Grid (4x5)	0.3	0.993
		Swipe	0.05	1
		Md. Grid (5x8)	-0.1	1
		Lg. Grid (8x10)	-0.75	0.614
		Zoom Grid (8x10)	-0.3	0.993
		RSVP	-0.05	1
	Lg. Grid (8x10)	Sm. Grid (4x5)	1.05	0.213
		Swipe	0.8	0.538
		Md. Grid (5x8)	0.65	0.758
		Sushi	0.75	0.614
		Zoom Grid (8x10)	0.45	0.948
		RSVP	0.7	0.688
	Zoom Grid (8x10)	Sm. Grid (4x5)	0.6	0.82
		Swipe	0.35	0.985
		Md. Grid (5x8)	0.2	0.999
		Sushi	0.3	0.993
		Lg. Grid (8x10)	-0.45	0.948
		RSVP	0.25	0.998
	RSVP	Sm. Grid (4x5)	0.35	0.985
		Swipe	0.1	1
		Md. Grid (5x8)	-0.05	1
		Sushi	0.05	1
		Lg. Grid (8x10)	-0.7	0.688
		Zoom Grid (8x10)	-0.25	0.998
Rushed	Sm. Grid (4x5)	Swipe	-0.25	0.994
		Md. Grid (5x8)	-0.1	1
		Sushi	-1.250*	0.017
		Lg. Grid (8x10)	-0.35	0.965
		Zoom Grid (8x10)	-0.35	0.965
		RSVP	-2.350*	0
	Swipe	Sm. Grid (4x5)	0.25	0.994
		Md. Grid (5x8)	0.15	1
		Sushi	-1	0.107
		Lg. Grid (8x10)	-0.1	1

		Zoom Grid (8x10)	-0.1	1
		RSVP	-2.100*	0
	Md. Grid (5x8)	Sm. Grid (4x5)	0.1	1
		Swipe	-0.15	1
		Sushi	-1.150*	0.037
		Lg. Grid (8x10)	-0.25	0.994
		Zoom Grid (8x10)	-0.25	0.994
		RSVP	-2.250*	0
	Sushi	Sm. Grid (4x5)	1.250*	0.017
		Swipe	1	0.107
		Md. Grid (5x8)	1.150*	0.037
		Lg. Grid (8x10)	0.9	0.196
		Zoom Grid (8x10)	0.9	0.196
		RSVP	-1.1	0.054
	Lg. Grid (8x10)	Sm. Grid (4x5)	0.35	0.965
		Swipe	0.1	1
		Md. Grid (5x8)	0.25	0.994
		Sushi	-0.9	0.196
		Zoom Grid (8x10)	0	1
		RSVP	-2.000*	0
	Zoom Grid (8x10)	Sm. Grid (4x5)	0.35	0.965
		Swipe	0.1	1
		Md. Grid (5x8)	0.25	0.994
		Sushi	-0.9	0.196
		Lg. Grid (8x10)	0	1
		RSVP	-2.000*	0
	RSVP	Sm. Grid (4x5)	2.350*	0
		Swipe	2.100*	0
		Md. Grid (5x8)	2.250*	0
		Sushi	1.1	0.054
		Lg. Grid (8x10)	2.000*	0
		Zoom Grid (8x10)	2.000*	0
Subjective Accuracy	Sm. Grid (4x5)	Swipe	0.45	0.813
		Md. Grid (5x8)	0.55	0.628
		Sushi	0.45	0.813
		Lg. Grid (8x10)	1.100*	0.017
		Zoom Grid (8x10)	0.35	0.935
		RSVP	1.050*	0.027
	Swipe	Sm. Grid (4x5)	-0.45	0.813
		Md. Grid (5x8)	0.1	1
		Sushi	0	1
		Lg. Grid (8x10)	0.65	0.426
		Zoom Grid (8x10)	-0.1	1
		RSVP	0.6	0.526
	Md. Grid (5x8)	Sm. Grid (4x5)	-0.55	0.628
		Swipe	-0.1	1
		Sushi	-0.1	1
		Lg. Grid (8x10)	0.55	0.628
		Zoom Grid (8x10)	-0.2	0.996
		RSVP	0.5	0.726
	Sushi	Sm. Grid (4x5)	-0.45	0.813
		Swipe	0	1
		Md. Grid (5x8)	0.1	1
		Lg. Grid (8x10)	0.65	0.426

		Zoom Grid (8x10)	-0.1	1
		RSVP	0.6	0.526
	Lg. Grid (8x10)	Sm. Grid (4x5)	-1.100*	0.017
		Swipe	-0.65	0.426
		Md. Grid (5x8)	-0.55	0.628
		Sushi	-0.65	0.426
		Zoom Grid (8x10)	-0.75	0.254
		RSVP	-0.05	1
	Zoom Grid (8x10)	Sm. Grid (4x5)	-0.35	0.935
		Swipe	0.1	1
		Md. Grid (5x8)	0.2	0.996
		Sushi	0.1	1
		Lg. Grid (8x10)	0.75	0.254
		RSVP	0.7	0.334
	RSVP	Sm. Grid (4x5)	-1.050*	0.027
		Swipe	-0.6	0.526
		Md. Grid (5x8)	-0.5	0.726
		Sushi	-0.6	0.526
		Lg. Grid (8x10)	0.05	1
		Zoom Grid (8x10)	-0.7	0.334
Visual Appeal	Sm. Grid (4x5)	Swipe	0.2	0.997
		Md. Grid (5x8)	0.65	0.448
		Sushi	0.35	0.94
		Lg. Grid (8x10)	1.150*	0.013
		Zoom Grid (8x10)	0.4	0.892
		RSVP	0.75	0.274
	Swipe	Sm. Grid (4x5)	-0.2	0.997
		Md. Grid (5x8)	0.45	0.825
		Sushi	0.15	0.999
		Lg. Grid (8x10)	0.95	0.072
		Zoom Grid (8x10)	0.2	0.997
		RSVP	0.55	0.648
	Md. Grid (5x8)	Sm. Grid (4x5)	-0.65	0.448
		Swipe	-0.45	0.825
		Sushi	-0.3	0.972
		Lg. Grid (8x10)	0.5	0.742
		Zoom Grid (8x10)	-0.25	0.989
		RSVP	0.1	1
	Sushi	Sm. Grid (4x5)	-0.35	0.94
		Swipe	-0.15	0.999
		Md. Grid (5x8)	0.3	0.972
		Lg. Grid (8x10)	0.8	0.204
		Zoom Grid (8x10)	0.05	1
		RSVP	0.4	0.892
	Lg. Grid (8x10)	Sm. Grid (4x5)	-1.150*	0.013
		Swipe	-0.95	0.072
		Md. Grid (5x8)	-0.5	0.742
		Sushi	-0.8	0.204
		Zoom Grid (8x10)	-0.75	0.274
		RSVP	-0.4	0.892
	Zoom Grid (8x10)	Sm. Grid (4x5)	-0.4	0.892
		Swipe	-0.2	0.997
		Md. Grid (5x8)	0.25	0.989
		Sushi	-0.05	1

		Lg. Grid (8x10)	0.75	0.274
		RSVP	0.35	0.94
	RSVP	Sm. Grid (4x5)	-0.75	0.274
		Swipe	-0.55	0.648
		Md. Grid (5x8)	-0.1	1
		Sushi	-0.4	0.892
		Lg. Grid (8x10)	0.4	0.892
		Zoom Grid (8x10)	-0.35	0.94
Mental Demand	Sm. Grid (4x5)	Swipe	-0.15	1
		Md. Grid (5x8)	-0.35	0.977
		Sushi	-0.7	0.6
		Lg. Grid (8x10)	-1.2	0.055
		Zoom Grid (8x10)	-0.55	0.824
		RSVP	-1.600*	0.002
	Swipe	Sm. Grid (4x5)	0.15	1
		Md. Grid (5x8)	-0.2	0.999
		Sushi	-0.55	0.824
		Lg. Grid (8x10)	-1.05	0.138
		Zoom Grid (8x10)	-0.4	0.956
		RSVP	-1.450*	0.009
	Md. Grid (5x8)	Sm. Grid (4x5)	0.35	0.977
		Swipe	0.2	0.999
		Sushi	-0.35	0.977
		Lg. Grid (8x10)	-0.85	0.361
		Zoom Grid (8x10)	-0.2	0.999
		RSVP	-1.250*	0.039
	Sushi	Sm. Grid (4x5)	0.7	0.6
		Swipe	0.55	0.824
		Md. Grid (5x8)	0.35	0.977
		Lg. Grid (8x10)	-0.5	0.88
		Zoom Grid (8x10)	0.15	1
		RSVP	-0.9	0.293
	Lg. Grid (8x10)	Sm. Grid (4x5)	1.2	0.055
		Swipe	1.05	0.138
		Md. Grid (5x8)	0.85	0.361
		Sushi	0.5	0.88
		Zoom Grid (8x10)	0.65	0.681
		RSVP	-0.4	0.956
	Zoom Grid (8x10)	Sm. Grid (4x5)	0.55	0.824
		Swipe	0.4	0.956
		Md. Grid (5x8)	0.2	0.999
		Sushi	-0.15	1
		Lg. Grid (8x10)	-0.65	0.681
		RSVP	-1.05	0.138
	RSVP	Sm. Grid (4x5)	1.600*	0.002
		Swipe	1.450*	0.009
		Md. Grid (5x8)	1.250*	0.039
		Sushi	0.9	0.293
		Lg. Grid (8x10)	0.4	0.956
		Zoom Grid (8x10)	1.05	0.138
Physical Demand	Sm. Grid (4x5)	Swipe	-0.1	1
		Md. Grid (5x8)	0.05	1
		Sushi	-0.15	1
		Lg. Grid (8x10)	-0.6	0.675

		Zoom Grid (8x10)	-0.35	0.965
		RSVP	-0.7	0.497
	Swipe	Sm. Grid (4x5)	0.1	1
		Md. Grid (5x8)	0.15	1
		Sushi	-0.05	1
		Lg. Grid (8x10)	-0.5	0.83
		Zoom Grid (8x10)	-0.25	0.994
		RSVP	-0.6	0.675
	Md. Grid (5x8)	Sm. Grid (4x5)	-0.05	1
		Swipe	-0.15	1
		Sushi	-0.2	0.998
		Lg. Grid (8x10)	-0.65	0.587
		Zoom Grid (8x10)	-0.4	0.935
		RSVP	-0.75	0.411
	Sushi	Sm. Grid (4x5)	0.15	1
		Swipe	0.05	1
		Md. Grid (5x8)	0.2	0.998
		Lg. Grid (8x10)	-0.45	0.89
		Zoom Grid (8x10)	-0.2	0.998
		RSVP	-0.55	0.758
	Lg. Grid (8x10)	Sm. Grid (4x5)	0.6	0.675
		Swipe	0.5	0.83
		Md. Grid (5x8)	0.65	0.587
		Sushi	0.45	0.89
		Zoom Grid (8x10)	0.25	0.994
		RSVP	-0.1	1
	Zoom Grid (8x10)	Sm. Grid (4x5)	0.35	0.965
		Swipe	0.25	0.994
		Md. Grid (5x8)	0.4	0.935
		Sushi	0.2	0.998
		Lg. Grid (8x10)	-0.25	0.994
		RSVP	-0.35	0.965
	RSVP	Sm.	0.7	0.497
		Swipe	0.6	0.675
		Md. Grid (5x8)	0.75	0.411
		Sushi	0.55	0.758
		Lg. Grid (8x10)	0.1	1
		Zoom Grid (8x10)	0.35	0.965

Appendix F

ANOVA Results - Study 2

Table F.1: Accuracy ANOVA results for Study 2. Significant results are starred.

		Sum of Squares	df	Mean Square	F	Sig.
Correct	Between Groups	384.222	2	192.111	1.209	0.316
	Within Groups	3814.444	24	158.935		
	Total	4198.667	26			
Precision	Between Groups	0.017	2	0.009	0.326	0.725
	Within Groups	0.643	24	0.027		
	Total	0.66	26			
Recall	Between Groups	0.015	2	0.008	0.525	0.598
	Within Groups	0.345	24	0.014		
	Total	0.36	26			

Table F.2: TLX ANOVA results for Study 2. Significant results are starred.

		Sum of Squares	df	Mean Square	F	Sig.
Easy to Use	Between Groups	12.963	2	6.481	4.348	0.024*
	Within Groups	35.778	24	1.491		
	Total	48.741	26			
Easy Decisions	Between Groups	18.667	2	9.333	11.2	0*
	Within Groups	20	24	0.833		
	Total	38.667	26			
Overwhelmed	Between Groups	2.667	2	1.333	1.143	0.336
	Within Groups	28	24	1.167		
	Total	30.667	26			
Rushed	Between Groups	30.889	2	15.444	18.742	0*
	Within Groups	19.778	24	0.824		
	Total	50.667	26			
Subjective Accuracy	Between Groups	9.556	2	4.778	5.432	0.011*
	Within Groups	21.111	24	0.88		
	Total	30.667	26			
Visual Appeal	Between Groups	2.074	2	1.037	1.087	0.353
	Within Groups	22.889	24	0.954		
	Total	24.963	26			
Mental Demand	Between Groups	8.074	2	4.037	2.595	0.095
	Within Groups	37.333	24	1.556		
	Total	45.407	26			
Physical Demand	Between Groups	1.407	2	0.704	0.422	0.66
	Within Groups	40	24	1.667		
	Total	41.407	26			

Appendix G

Tukey Results - Study 2

Table G.1: Accuracy Tukey HSD results for Study 2. Significant results are starred.

Dependent Variable	Interface Style (I)	Interface Style (J)	(I-J) Mean Difference	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Correct	Slow	Med	5.111	5.943	0.67	-9.73	19.95
		Fast	9.222	5.943	0.286	-5.62	24.06
	Med	Slow	-5.111	5.943	0.67	-19.95	9.73
		Fast	4.111	5.943	0.771	-10.73	18.95
	Fast	Slow	-9.222	5.943	0.286	-24.06	5.62
		Med	-4.111	5.943	0.771	-18.95	10.73
Precision	Slow	Med	0.03318	0.07714	0.903	-0.15945	0.22582
		Fast	0.06220	0.07714	0.703	-0.13044	0.25483
	Med	Slow	-0.03318	0.07714	0.903	-0.22582	0.15945
		Fast	0.02902	0.07714	0.925	-0.16362	0.22165
	Fast	Slow	-0.06220	0.07714	0.703	-0.25484	0.13044
		Med	-0.02902	0.07714	0.925	-0.22165	0.16362
Recall	Slow	Med	0.00694	0.05651	0.992	-0.13418	0.14807
		Fast	0.05324	0.05651	0.62	-0.08789	0.19437
	Med	Slow	-0.00694	0.05651	0.992	-0.14807	0.13418
		Fast	0.04630	0.05651	0.695	-0.09483	0.18742
	Fast	Slow	-0.05324	0.05651	0.62	-0.194367	0.08789
		Med	-0.046296	0.05651	0.695	-0.18742	0.09483

Table G.2: TLX Tukey HSD results for Study 2. Significant results are starred.

Dependent Variable	Interface Style (I)	Interface Style (J)	(I-J) Mean Difference	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Easy to Use	Slow	Med	0.55556	0.57557	0.605	-0.8818	1.9929
		Fast	1.66667*	0.57557	0.021	0.2293	3.104
	Med	Slow	-0.55556	0.57557	0.605	-1.9929	0.8818
		Fast	1.11111	0.57557	0.152	-0.3262	2.5485
	Fast	Slow	-1.66667*	0.57557	0.021	-3.104	-0.2293
		Med	-1.11111	0.57557	0.152	-2.5485	0.3262
Easy Decisions	Slow	Med	1.33333*	0.43033	0.013	0.2587	2.408
		Fast	2.00000*	0.43033	0	0.9253	3.0747
	Med	Slow	-1.33333*	0.43033	0.013	-2.408	-0.2587
		Fast	0.66667	0.43033	0.287	-0.408	1.7413
	Fast	Slow	-2.00000*	0.43033	0	-3.0747	-0.9253
		Med	-0.66667	0.43033	0.287	-1.7413	0.408

Overwhelmed	Slow	Med	0	0.50918	1	-1.2716	1.2716
		Fast	-0.66667	0.50918	0.404	-1.9382	0.6049
	Med	Slow	0	0.50918	1	-1.2716	1.2716
		Fast	-0.66667	0.50918	0.404	-1.9382	0.6049
	Fast	Slow	0.66667	0.50918	0.404	-0.6049	1.9382
		Med	0.66667	0.50918	0.404	-0.6049	1.9382
Rushed	Slow	Med	-0.77778	0.42793	0.185	-1.8465	0.2909
		Fast	-2.55556*	0.42793	0	-3.6242	-1.4869
	Med	Slow	0.77778	0.42793	0.185	-0.2909	1.8465
		Fast	-1.77778*	0.42793	0.001	-2.8465	-0.7091
	Fast	Slow	2.55556*	0.42793	0	1.4869	3.6242
		Med	1.77778*	0.42793	0.001	0.7091	2.8465
Subjective Accuracy	Slow	Med	0.55556	0.44212	0.433	-0.5486	1.6597
		Fast	1.44444*	0.44212	0.009	0.3403	2.5486
	Med	Slow	-0.55556	0.44212	0.433	-1.6597	0.5486
		Fast	0.88889	0.44212	0.131	-0.2152	1.993
	Fast	Slow	-1.44444*	0.44212	0.009	-2.5486	-0.3403
		Med	-0.88889	0.44212	0.131	-1.993	0.2152
Visual Appeal	Slow	Med	0.22222	0.46036	0.88	-0.9274	1.3719
		Fast	0.66667	0.46036	0.333	-0.483	1.8163
	Med	Slow	-0.22222	0.46036	0.88	-1.3719	0.9274
		Fast	0.44444	0.46036	0.605	-0.7052	1.5941
	Fast	Slow	-0.66667	0.46036	0.333	-1.8163	0.483
		Med	-0.44444	0.46036	0.605	-1.5941	0.7052
Mental Demand	Slow	Med	-0.77778	0.58794	0.397	-2.246	0.6905
		Fast	-1.33333	0.58794	0.08	-2.8016	0.1349
	Med	Slow	0.77778	0.58794	0.397	-0.6905	2.246
		Fast	-0.55556	0.58794	0.618	-2.0238	0.9127
	Fast	Slow	1.33333	0.58794	0.08	-0.1349	2.8016
		Med	0.55556	0.58794	0.618	-0.9127	2.0238
Physical Demand	Slow	Med	-0.22222	0.60858	0.929	-1.742	1.2976
		Fast	-0.55556	0.60858	0.638	-2.0754	0.9642
	Med	Slow	0.22222	0.60858	0.929	-1.2976	1.742
		Fast	-0.33333	0.60858	0.849	-1.8531	1.1865
	Fast	Slow	0.55556	0.60858	0.638	-0.9642	2.0754
		Med	0.33333	0.60858	0.849	-1.1865	1.8531

Appendix H

ANOVA Results - Study 3

Table H.1: Time ANOVA results for Study 3. Significant results are starred.

		Sum of Squares	df	Mean Square	F	Sig.
Time	Between Groups	1307871.951	5	261574.39	24.606	0*
	Within Groups	1084303.496	102	10630.426		
	Total	2392175.448	107			

Table H.2: Accuracy ANOVA results for Study 3. Significant results are starred.

		Sum of Squares	df	Mean Square	F	Sig.
Correct Responses	Between Groups	1172.491	5	234.498	0.344	0.885
	Within Groups	69584.056	102	682.197		
	Total	70756.546	107			
Precision	Between Groups	0.125	5	0.025	0.794	0.557
	Within Groups	3.219	102	0.032		
	Total	3.345	107			
Recall	Between Groups	0.503	5	0.101	8.3	0*
	Within Groups	1.237	102	0.012		
	Total	1.74	107			

Table H.3: TLX ANOVA results for Study 3. Significant results are starred.

		Sum of Squares	df	Mean Square	F	Sig.
Easy To Use	Between Groups	54.407	5	10.881	7.388	0*
	Within Groups	150.222	102	1.473		
	Total	204.63	107			
Easy Decisions	Between Groups	36.157	5	7.231	4.561	0.001*
	Within Groups	161.722	102	1.586		
	Total	197.88	107			
Overwhelmed	Between Groups	40.185	5	8.037	5.568	0*
	Within Groups	147.222	102	1.443		
	Total	187.407	107			
Rushed	Between Groups	112.38	5	22.476	17.404	0*
	Within Groups	131.722	102	1.291		
	Total	244.102	107			
Subjective Accuracy	Between Groups	32.296	5	6.459	4.782	0.001*
	Within Groups	137.778	102	1.351		
	Total	170.074	107			
Visual Appeal	Between Groups	26.074	5	5.215	3.728	0.004*
	Within Groups	142.667	102	1.399		
	Total	168.741	107			
Mental Demand	Between Groups	30.444	5	6.089	4.493	0.001*
	Within Groups	138.222	102	1.355		
	Total	168.667	107			
Physical Demand	Between Groups	6.074	5	1.215	0.669	0.648

	Within Groups	185.111	102	1.815		
	Total	191.185	107			

Appendix I

Tukey Results - Study 3

Table I.1: Time Tukey HSD results for Study 3. Significant results are starred.

Dependent Variable	Interface Style (I)	Interface Style (J)	(I-J) Mean Difference	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Time	Sm. Grid	Swipe	-229.3107	34.3680	0*	-329.1364	-129.4849
		Sushi	-215.1806	34.3680	0*	-315.0064	-115.3548
		Lg. Grid	21.9586	34.3680	0.988	-77.8672	121.7844
		RSVP	-248.2158	34.3680	0*	-348.0416	-148.3900
		Preview Swipe	-178.7939	34.3680	0*	-278.6196	-78.9681
	Swipe	Sm. Grid	229.3107	34.3680	0*	129.4849	329.1364
		Sushi	14.1300	34.3680	0.998	-85.6957	113.9558
		Lg. Grid	251.2693	34.3680	0*	151.4435	351.0950
		RSVP	-18.9051	34.3680	0.994	-118.7309	80.9206
		Preview Swipe	50.5168	34.3680	0.684	-49.3090	150.3426
	Sushi	Sm. Grid	215.1806	34.3680	0*	115.3548	315.0064
		Swipe	-14.1300	34.3680	0.998	-113.9558	85.6957
		Lg. Grid	237.1392	34.3680	0*	137.3134	336.9650
		RSVP	-33.0352	34.3680	0.929	-132.8610	66.7906
		Preview Swipe	36.3868	34.3680	0.896	-63.4390	136.2126
	Lg. Grid	Sm. Grid	-21.9586	34.3680	0.988	-121.7844	77.8672
		Swipe	-251.2693	34.3680	0*	-351.0950	-151.4435
		Sushi	-237.1392	34.3680	0*	-336.9650	-137.3134
		RSVP	-270.1744	34.3680	0*	-370.0002	-170.3486
		Preview Swipe	-200.7524	34.3680	0*	-300.5782	-100.9267
	RSVP	Sm. Grid	248.2158	34.3680	0*	148.3900	348.0416
		Swipe	18.9051	34.3680	0.994	-80.9206	118.7309
		Sushi	33.0352	34.3680	0.929	-66.7906	132.8610
		Lg. Grid	270.1744	34.3680	0*	170.3486	370.0002
		Preview Swipe	69.4219	34.3680	0.338	-30.4038	169.2477
	Preview Swipe	Sm. Grid	178.7939	34.3680	0	78.9681	278.6196
		Swipe	-50.5168	34.3680	0.684	-150.3426	49.3090
		Sushi	-36.3868	34.3680	0.896	-136.2126	63.4390
		Lg. Grid	200.7524	34.3680	0*	100.9267	300.5782
		RSVP	-69.4219	34.3680	0.338	-169.2477	30.4038

Table I.2: Accuracy Tukey HSD results for Study 3. Significant results are starred.

Dependent Variable	Interface Style (I)	Interface Style (J)	(I-J) Mean Difference	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Correct Responses	Sm. Grid	Swipe	-2	8.706	1	-27.29	23.29
		Sushi	0.778	8.706	1	-24.51	26.07
		Lg. Grid	3.111	8.706	0.999	-22.18	28.4
		RSVP	-7.444	8.706	0.956	-32.73	17.84
		Preview Swipe	-2.5	8.706	1	-27.79	22.79
	Swipe	Sm. Grid	2	8.706	1	-23.29	27.29
		Sushi	2.778	8.706	1	-22.51	28.07
		Lg. Grid	5.111	8.706	0.992	-20.18	30.4
		RSVP	-5.444	8.706	0.989	-30.73	19.84
		Preview Swipe	-0.5	8.706	1	-25.79	24.79
	Sushi	Sm. Grid	-0.778	8.706	1	-26.07	24.51
		Swipe	-2.778	8.706	1	-28.07	22.51
		Lg. Grid	2.333	8.706	1	-22.96	27.62
		RSVP	-8.222	8.706	0.934	-33.51	17.07
		Preview Swipe	-3.278	8.706	0.999	-28.57	22.01
	Lg. Grid	Sm. Grid	-3.111	8.706	0.999	-28.4	22.18
		Swipe	-5.111	8.706	0.992	-30.4	20.18
		Sushi	-2.333	8.706	1	-27.62	22.96
		RSVP	-10.556	8.706	0.83	-35.84	14.73
		Preview Swipe	-5.611	8.706	0.987	-30.9	19.68
	RSVP	Sm. Grid	7.444	8.706	0.956	-17.84	32.73
		Swipe	5.444	8.706	0.989	-19.84	30.73
		Sushi	8.222	8.706	0.934	-17.07	33.51
		Lg. Grid	10.556	8.706	0.83	-14.73	35.84
		Preview Swipe	4.944	8.706	0.993	-20.34	30.23
	Preview Swipe	Sm. Grid	2.5	8.706	1	-22.79	27.79
		Swipe	0.5	8.706	1	-24.79	25.79
		Sushi	3.278	8.706	0.999	-22.01	28.57
		Lg. Grid	5.611	8.706	0.987	-19.68	30.9
		RSVP	-4.944	8.706	0.993	-30.23	20.34
Precision	Sm. Grid	Swipe	2.35E-03	5.92E-02	1	-1.70E-01	1.74E-01
		Sushi	3.79E-02	5.92E-02	0.988	-1.34E-01	2.10E-01
		Lg. Grid	-2.60E-02	5.92E-02	0.998	-1.98E-01	1.46E-01
		RSVP	5.70E-02	5.92E-02	0.928	-1.15E-01	2.29E-01
		Preview Swipe	7.02E-02	5.92E-02	0.843	-1.02E-01	2.42E-01
	Swipe	Sm. Grid	-2.35E-03	5.92E-02	1	-1.74E-01	1.70E-01
		Sushi	3.55E-02	5.92E-02	0.991	-1.36E-01	2.08E-01
		Lg. Grid	-2.84E-02	5.92E-02	0.997	-2.00E-01	1.44E-01
		RSVP	5.47E-02	5.92E-02	0.94	-1.17E-01	2.27E-01
		Preview Swipe	6.78E-02	5.92E-02	0.861	-1.04E-01	2.40E-01

	Sushi	Sm. Grid	-3.79E-02	5.92E-02	0.988	-2.10E-01	1.34E-01
		Swipe	-3.55E-02	5.92E-02	0.991	-2.08E-01	1.36E-01
		Lg. Grid	-6.39E-02	5.92E-02	0.889	-2.36E-01	1.08E-01
		RSVP	1.92E-02	5.92E-02	1	-1.53E-01	1.91E-01
		Preview Swipe	3.23E-02	5.92E-02	0.994	-1.40E-01	2.04E-01
	Lg. Grid	Sm. Grid	2.60E-02	5.92E-02	0.998	-1.46E-01	1.98E-01
		Swipe	2.84E-02	5.92E-02	0.997	-1.44E-01	2.00E-01
		Sushi	6.39E-02	5.92E-02	0.889	-1.08E-01	2.36E-01
		RSVP	8.30E-02	5.92E-02	0.726	-8.90E-02	2.55E-01
		Preview Swipe	9.62E-02	5.92E-02	0.585	-7.58E-02	2.68E-01
	RSVP	Sm. Grid	-5.70E-02	5.92E-02	0.928	-2.29E-01	1.15E-01
		Swipe	-5.47E-02	5.92E-02	0.94	-2.27E-01	1.17E-01
		Sushi	-1.92E-02	5.92E-02	1	-1.91E-01	1.53E-01
		Lg. Grid	-8.30E-02	5.92E-02	0.726	-2.55E-01	8.90E-02
		Preview Swipe	1.31E-02	5.92E-02	1	-1.59E-01	1.85E-01
	Preview Swipe	Sm. Grid	-7.02E-02	5.92E-02	0.843	-2.42E-01	1.02E-01
		Swipe	-6.78E-02	5.92E-02	0.861	-2.40E-01	1.04E-01
		Sushi	-3.23E-02	5.92E-02	0.994	-2.04E-01	1.40E-01
		Lg. Grid	-9.62E-02	5.92E-02	0.585	-2.68E-01	7.58E-02
		RSVP	-1.31E-02	5.92E-02	1	-1.85E-01	1.59E-01
Recall	Sm. Grid	Swipe	-4.28E-02	3.67E-02	0.851	-1.49E-01	6.38E-02
		Sushi	-8.91E-02	3.67E-02	0.156	-1.96E-01	1.75E-02
		Lg. Grid	1.17E-01*	3.67E-02	0.023	1.03E-02	2.24E-01
		RSVP	-6.45E-02	3.67E-02	0.497	-1.71E-01	4.21E-02
		Preview Swipe	-5.99E-02	3.67E-02	0.58	-1.67E-01	4.67E-02
	Swipe	Sm. Grid	4.28E-02	3.67E-02	0.851	-6.38E-02	1.49E-01
		Sushi	-4.63E-02	3.67E-02	0.805	-1.53E-01	6.03E-02
		Lg. Grid	1.60E-01*	3.67E-02	0	5.31E-02	2.66E-01
		RSVP	-2.17E-02	3.67E-02	0.991	-1.28E-01	8.49E-02
		Preview Swipe	-1.71E-02	3.67E-02	0.997	-1.24E-01	8.95E-02
	Sushi	Sm. Grid	8.91E-02	3.67E-02	0.156	-1.75E-02	1.96E-01
		Swipe	4.63E-02	3.67E-02	0.805	-6.03E-02	1.53E-01
		Lg. Grid	2.06E-01*	3.67E-02	0	9.94E-02	3.13E-01
		RSVP	2.46E-02	3.67E-02	0.985	-8.20E-02	1.31E-01
		Preview Swipe	2.92E-02	3.67E-02	0.968	-7.74E-02	1.36E-01
	Lg. Grid	Sm. Grid	-1.17E-01*	3.67E-02	0.023	-2.24E-01	-1.03E-02
		Swipe	-1.60E-01*	3.67E-02	0	-2.66E-01	-5.31E-02
		Sushi	-2.06E-01*	3.67E-02	0	-3.13E-01	-9.94E-02
		RSVP	-1.81E-01*	3.67E-02	0	-2.88E-01	-7.48E-02
		Preview Swipe	-1.77E-01*	3.67E-02	0	-2.83E-01	-7.02E-02
	RSVP	Sm. Grid	6.45E-02	3.67E-02	0.497	-4.21E-02	1.71E-01
		Swipe	2.17E-02	3.67E-02	0.991	-8.49E-02	1.28E-01
		Sushi	-2.46E-02	3.67E-02	0.985	-1.31E-01	8.20E-02
		Lg. Grid	1.81E-01*	3.67E-02	0	7.48E-02	2.88E-01
		Preview Swipe	4.63E-03	3.67E-02	1	-1.02E-01	1.11E-01

	Preview Swipe	Sm. Grid	5.99E-02	3.67E-02	0.58	-4.67E-02	1.67E-01
		Swipe	1.71E-02	3.67E-02	0.997	-8.95E-02	1.24E-01
		Sushi	-2.92E-02	3.67E-02	0.968	-1.36E-01	7.74E-02
		Lg. Grid	1.77E-01*	3.67E-02	0	7.02E-02	2.83E-01
		RSVP	-4.63E-03	3.67E-02	1	-1.11E-01	1.02E-01

Table I.3: TLX Tukey HSD results for Study 3. Significant results are starred.

Dependent Variable	Interface Style (I)	Interface Style (J)	(I-J) Mean Difference	Std. Error	Sig.	95% Confidence Interval Lower Bound	Upper Bound
Easy to Use	Sm. Grid	Swipe	-0.833	0.405	0.317	-2.01	0.34
		Sushi	0.611	0.405	0.658	-0.56	1.79
		Lg. Grid	1.278*	0.405	0.025	0.1	2.45
		RSVP	0.611	0.405	0.658	-0.56	1.79
		Preview Swipe	-0.444	0.405	0.881	-1.62	0.73
	Swipe	Sm. Grid	0.833	0.405	0.317	-0.34	2.01
		Sushi	1.444*	0.405	0.007	0.27	2.62
		Lg. Grid	2.111*	0.405	0	0.94	3.29
		RSVP	1.444*	0.405	0.007	0.27	2.62
		Preview Swipe	0.389	0.405	0.929	-0.79	1.56
	Sushi	Sm. Grid	-0.611	0.405	0.658	-1.79	0.56
		Swipe	-1.444*	0.405	0.007	-2.62	-0.27
		Lg. Grid	0.667	0.405	0.569	-0.51	1.84
		RSVP	0	0.405	1	-1.17	1.17
		Preview Swipe	-1.056	0.405	0.104	-2.23	0.12
	Lg. Grid	Sm. Grid	-1.278*	0.405	0.025	-2.45	-0.1
		Swipe	-2.111*	0.405	0	-3.29	-0.94
		Sushi	-0.667	0.405	0.569	-1.84	0.51
		RSVP	-0.667	0.405	0.569	-1.84	0.51
		Preview Swipe	-1.722*	0.405	0.001	-2.9	-0.55
	RSVP	Sm. Grid	-0.611	0.405	0.658	-1.79	0.56
		Swipe	-1.444*	0.405	0.007	-2.62	-0.27
		Sushi	0	0.405	1	-1.17	1.17
		Lg. Grid	0.667	0.405	0.569	-0.51	1.84
		Preview Swipe	-1.056	0.405	0.104	-2.23	0.12
	Preview Swipe	Sm. Grid	0.444	0.405	0.881	-0.73	1.62
		Swipe	-0.389	0.405	0.929	-1.56	0.79
		Sushi	1.056	0.405	0.104	-0.12	2.23
		Lg. Grid	1.722*	0.405	0.001	0.55	2.9
		RSVP	1.056	0.405	0.104	-0.12	2.23
Easy Decisions	Sm. Grid	Swipe	-1	0.42	0.172	-2.22	0.22
		Sushi	-0.167	0.42	0.999	-1.39	1.05

		Lg. Grid	0.611	0.42	0.693	-0.61	1.83
		RSVP	0.5	0.42	0.84	-0.72	1.72
		Preview Swipe	-0.667	0.42	0.608	-1.89	0.55
	Swipe	Sm. Grid	1	0.42	0.172	-0.22	2.22
		Sushi	0.833	0.42	0.358	-0.39	2.05
		Lg. Grid	1.611*	0.42	0.003	0.39	2.83
		RSVP	1.500*	0.42	0.007	0.28	2.72
		Preview Swipe	0.333	0.42	0.968	-0.89	1.55
	Sushi	Sm. Grid	0.167	0.42	0.999	-1.05	1.39
		Swipe	-0.833	0.42	0.358	-2.05	0.39
		Lg. Grid	0.778	0.42	0.437	-0.44	2
		RSVP	0.667	0.42	0.608	-0.55	1.89
		Preview Swipe	-0.5	0.42	0.84	-1.72	0.72
	Lg. Grid	Sm. Grid	-0.611	0.42	0.693	-1.83	0.61
		Swipe	-1.611*	0.42	0.003	-2.83	-0.39
		Sushi	-0.778	0.42	0.437	-2	0.44
		RSVP	-0.111	0.42	1	-1.33	1.11
		Preview Swipe	-1.278*	0.42	0.034	-2.5	-0.06
	RSVP	Sm. Grid	-0.5	0.42	0.84	-1.72	0.72
		Swipe	-1.500*	0.42	0.007	-2.72	-0.28
		Sushi	-0.667	0.42	0.608	-1.89	0.55
		Lg. Grid	0.111	0.42	1	-1.11	1.33
		Preview Swipe	-1.167	0.42	0.069	-2.39	0.05
	Preview Swipe	Sm. Grid	0.667	0.42	0.608	-0.55	1.89
		Swipe	-0.333	0.42	0.968	-1.55	0.89
		Sushi	0.5	0.42	0.84	-0.72	1.72
		Lg. Grid	1.278*	0.42	0.034	0.06	2.5
		RSVP	1.167	0.42	0.069	-0.05	2.39
Overwhelmed	Sm. Grid	Swipe	0.611	0.4	0.648	-0.55	1.77
		Sushi	-0.944	0.4	0.181	-2.11	0.22
		Lg. Grid	-0.944	0.4	0.181	-2.11	0.22
		RSVP	-1.056	0.4	0.098	-2.22	0.11
		Preview Swipe	-0.222	0.4	0.994	-1.39	0.94
	Swipe	Sm. Grid	-0.611	0.4	0.648	-1.77	0.55
		Sushi	-1.556*	0.4	0.002	-2.72	-0.39
		Lg. Grid	-1.556*	0.4	0.002	-2.72	-0.39
		RSVP	-1.667*	0.4	0.001	-2.83	-0.5
		Preview Swipe	-0.833	0.4	0.306	-2	0.33
	Sushi	Sm. Grid	0.944	0.4	0.181	-0.22	2.11
		Swipe	1.556*	0.4	0.002	0.39	2.72
		Lg. Grid	0	0.4	1	-1.16	1.16
		RSVP	-0.111	0.4	1	-1.27	1.05
		Preview Swipe	0.722	0.4	0.468	-0.44	1.89
	Lg. Grid	Sm. Grid	0.944	0.4	0.181	-0.22	2.11

		Swipe	1.556*	0.4	0.002	0.39	2.72
		Sushi	0	0.4	1	-1.16	1.16
		RSVP	-0.111	0.4	1	-1.27	1.05
		Preview Swipe	0.722	0.4	0.468	-0.44	1.89
	RSVP	Sm. Grid	1.056	0.4	0.098	-0.11	2.22
		Swipe	1.667*	0.4	0.001	0.5	2.83
		Sushi	0.111	0.4	1	-1.05	1.27
		Lg. Grid	0.111	0.4	1	-1.05	1.27
		Preview Swipe	0.833	0.4	0.306	-0.33	2
	Preview Swipe	Sm. Grid	0.222	0.4	0.994	-0.94	1.39
		Swipe	0.833	0.4	0.306	-0.33	2
		Sushi	-0.722	0.4	0.468	-1.89	0.44
		Lg. Grid	-0.722	0.4	0.468	-1.89	0.44
		RSVP	-0.833	0.4	0.306	-2	0.33
Rushed	Sm. Grid	Swipe	-0.722	0.379	0.404	-1.82	0.38
		Sushi	-1.722*	0.379	0	-2.82	-0.62
		Lg. Grid	0	0.379	1	-1.1	1.1
		RSVP	-2.611*	0.379	0	-3.71	-1.51
		Preview Swipe	0.111	0.379	1	-0.99	1.21
	Swipe	Sm. Grid	0.722	0.379	0.404	-0.38	1.82
		Sushi	-1	0.379	0.097	-2.1	0.1
		Lg. Grid	0.722	0.379	0.404	-0.38	1.82
		RSVP	-1.889*	0.379	0	-2.99	-0.79
		Preview Swipe	0.833	0.379	0.247	-0.27	1.93
	Sushi	Sm. Grid	1.722*	0.379	0	0.62	2.82
		Swipe	1	0.379	0.097	-0.1	2.1
		Lg. Grid	1.722*	0.379	0	0.62	2.82
		RSVP	-0.889	0.379	0.185	-1.99	0.21
		Preview Swipe	1.833*	0.379	0	0.73	2.93
	Lg. Grid	Sm. Grid	0	0.379	1	-1.1	1.1
		Swipe	-0.722	0.379	0.404	-1.82	0.38
		Sushi	-1.722*	0.379	0	-2.82	-0.62
		RSVP	-2.611*	0.379	0	-3.71	-1.51
		Preview Swipe	0.111	0.379	1	-0.99	1.21
	RSVP	Sm. Grid	2.611*	0.379	0	1.51	3.71
		Swipe	1.889*	0.379	0	0.79	2.99
		Sushi	0.889	0.379	0.185	-0.21	1.99
		Lg. Grid	2.611*	0.379	0	1.51	3.71
		Preview Swipe	2.722*	0.379	0	1.62	3.82
	Preview Swipe	Sm. Grid	-0.111	0.379	1	-1.21	0.99
		Swipe	-0.833	0.379	0.247	-1.93	0.27
		Sushi	-1.833*	0.379	0	-2.93	-0.73
		Lg. Grid	-0.111	0.379	1	-1.21	0.99
		RSVP	-2.722*	0.379	0	-3.82	-1.62

Subjective Accuracy	Sm. Grid	Swipe	-0.333	0.387	0.955	-1.46	0.79
		Sushi	0.111	0.387	1	-1.01	1.24
		Lg. Grid	0.889	0.387	0.206	-0.24	2.01
		RSVP	1.111	0.387	0.055	-0.01	2.24
		Preview Swipe	-0.222	0.387	0.993	-1.35	0.9
	Swipe	Sm. Grid	0.333	0.387	0.955	-0.79	1.46
		Sushi	0.444	0.387	0.86	-0.68	1.57
		Lg. Grid	1.222*	0.387	0.025	0.1	2.35
		RSVP	1.444*	0.387	0.004	0.32	2.57
		Preview Swipe	0.111	0.387	1	-1.01	1.24
	Sushi	Sm. Grid	-0.111	0.387	1	-1.24	1.01
		Swipe	-0.444	0.387	0.86	-1.57	0.68
		Lg. Grid	0.778	0.387	0.345	-0.35	1.9
		RSVP	1	0.387	0.111	-0.13	2.13
		Preview Swipe	-0.333	0.387	0.955	-1.46	0.79
	Lg. Grid	Sm. Grid	-0.889	0.387	0.206	-2.01	0.24
		Swipe	-1.222*	0.387	0.025	-2.35	-0.1
		Sushi	-0.778	0.387	0.345	-1.9	0.35
		RSVP	0.222	0.387	0.993	-0.9	1.35
		Preview Swipe	-1.111	0.387	0.055	-2.24	0.01
	RSVP	Sm. Grid	-1.111	0.387	0.055	-2.24	0.01
		Swipe	-1.444*	0.387	0.004	-2.57	-0.32
		Sushi	-1	0.387	0.111	-2.13	0.13
		Lg. Grid	-0.222	0.387	0.993	-1.35	0.9
		Preview Swipe	-1.333*	0.387	0.011	-2.46	-0.21
	Preview Swipe	Sm. Grid	0.222	0.387	0.993	-0.9	1.35
		Swipe	-0.111	0.387	1	-1.24	1.01
		Sushi	0.333	0.387	0.955	-0.79	1.46
		Lg. Grid	1.111	0.387	0.055	-0.01	2.24
		RSVP	1.333*	0.387	0.011	0.21	2.46
Visual Appeal	Sm. Grid	Swipe	-0.667	0.394	0.541	-1.81	0.48
		Sushi	0.111	0.394	1	-1.03	1.26
		Lg. Grid	1	0.394	0.123	-0.15	2.15
		RSVP	0.111	0.394	1	-1.03	1.26
		Preview Swipe	-0.111	0.394	1	-1.26	1.03
	Swipe	Sm. Grid	0.667	0.394	0.541	-0.48	1.81
		Sushi	0.778	0.394	0.365	-0.37	1.92
		Lg. Grid	1.667*	0.394	0.001	0.52	2.81
		RSVP	0.778	0.394	0.365	-0.37	1.92
		Preview Swipe	0.556	0.394	0.721	-0.59	1.7
	Sushi	Sm. Grid	-0.111	0.394	1	-1.26	1.03
		Swipe	-0.778	0.394	0.365	-1.92	0.37
		Lg. Grid	0.889	0.394	0.222	-0.26	2.03

		RSVP	0	0.394	1	-1.15	1.15
		Preview Swipe	-0.222	0.394	0.993	-1.37	0.92
	Lg. Grid	Sm. Grid	-1	0.394	0.123	-2.15	0.15
		Swipe	-1.667*	0.394	0.001	-2.81	-0.52
		Sushi	-0.889	0.394	0.222	-2.03	0.26
		RSVP	-0.889	0.394	0.222	-2.03	0.26
		Preview Swipe	-1.111	0.394	0.063	-2.26	0.03
	RSVP	Sm. Grid	-0.111	0.394	1	-1.26	1.03
		Swipe	-0.778	0.394	0.365	-1.92	0.37
		Sushi	0	0.394	1	-1.15	1.15
		Lg. Grid	0.889	0.394	0.222	-0.26	2.03
		Preview Swipe	-0.222	0.394	0.993	-1.37	0.92
	Preview Swipe	Sm. Grid	0.111	0.394	1	-1.03	1.26
		Swipe	-0.556	0.394	0.721	-1.7	0.59
		Sushi	0.222	0.394	0.993	-0.92	1.37
		Lg. Grid	1.111	0.394	0.063	-0.03	2.26
		RSVP	0.222	0.394	0.993	-0.92	1.37
Mental Demand	Sm. Grid	Swipe	0.444	0.388	0.861	-0.68	1.57
		Sushi	-0.722	0.388	0.432	-1.85	0.4
		Lg. Grid	-0.722	0.388	0.432	-1.85	0.4
		RSVP	-1.056	0.388	0.08	-2.18	0.07
		Preview Swipe	0.056	0.388	1	-1.07	1.18
	Swipe	Sm. Grid	-0.444	0.388	0.861	-1.57	0.68
		Sushi	-1.167*	0.388	0.038	-2.29	-0.04
		Lg. Grid	-1.167*	0.388	0.038	-2.29	-0.04
		RSVP	-1.500*	0.388	0.003	-2.63	-0.37
		Preview Swipe	-0.389	0.388	0.916	-1.52	0.74
	Sushi	Sm. Grid	0.722	0.388	0.432	-0.4	1.85
		Swipe	1.167*	0.388	0.038	0.04	2.29
		Lg. Grid	0	0.388	1	-1.13	1.13
		RSVP	-0.333	0.388	0.955	-1.46	0.79
		Preview Swipe	0.778	0.388	0.347	-0.35	1.9
	Lg. Grid	Sm. Grid	0.722	0.388	0.432	-0.4	1.85
		Swipe	1.167*	0.388	0.038	0.04	2.29
		Sushi	0	0.388	1	-1.13	1.13
		RSVP	-0.333	0.388	0.955	-1.46	0.79
		Preview Swipe	0.778	0.388	0.347	-0.35	1.9
	RSVP	Sm. Grid	1.056	0.388	0.08	-0.07	2.18
		Swipe	1.500*	0.388	0.003	0.37	2.63
		Sushi	0.333	0.388	0.955	-0.79	1.46
		Lg. Grid	0.333	0.388	0.955	-0.79	1.46
		Preview Swipe	1.111	0.388	0.056	-0.02	2.24
	Preview Swipe	Sm. Grid	-0.056	0.388	1	-1.18	1.07

		Swipe	0.389	0.388	0.916	-0.74	1.52
		Sushi	-0.778	0.388	0.347	-1.9	0.35
		Lg. Grid	-0.778	0.388	0.347	-1.9	0.35
		RSVP	-1.111	0.388	0.056	-2.24	0.02
Physical Demand	Sm. Grid	Swipe	-0.222	0.449	0.996	-1.53	1.08
		Sushi	-0.444	0.449	0.92	-1.75	0.86
		Lg. Grid	-0.611	0.449	0.75	-1.92	0.69
		RSVP	-0.556	0.449	0.818	-1.86	0.75
		Preview Swipe	-0.056	0.449	1	-1.36	1.25
	Swipe	Sm. Grid	0.222	0.449	0.996	-1.08	1.53
		Sushi	-0.222	0.449	0.996	-1.53	1.08
		Lg. Grid	-0.389	0.449	0.954	-1.69	0.92
		RSVP	-0.333	0.449	0.976	-1.64	0.97
		Preview Swipe	0.167	0.449	0.999	-1.14	1.47
	Sushi	Sm. Grid	0.444	0.449	0.92	-0.86	1.75
		Swipe	0.222	0.449	0.996	-1.08	1.53
		Lg. Grid	-0.167	0.449	0.999	-1.47	1.14
		RSVP	-0.111	0.449	1	-1.42	1.19
		Preview Swipe	0.389	0.449	0.954	-0.92	1.69
	Lg. Grid	Sm. Grid	0.611	0.449	0.75	-0.69	1.92
		Swipe	0.389	0.449	0.954	-0.92	1.69
		Sushi	0.167	0.449	0.999	-1.14	1.47
		RSVP	0.056	0.449	1	-1.25	1.36
		Preview Swipe	0.556	0.449	0.818	-0.75	1.86
	RSVP	Sm. Grid	0.556	0.449	0.818	-0.75	1.86
		Swipe	0.333	0.449	0.976	-0.97	1.64
		Sushi	0.111	0.449	1	-1.19	1.42
		Lg. Grid	-0.056	0.449	1	-1.36	1.25
		Preview Swipe	0.5	0.449	0.875	-0.8	1.8
	Preview Swipe	Sm. Grid	0.056	0.449	1	-1.25	1.36
		Swipe	-0.167	0.449	0.999	-1.47	1.14
		Sushi	-0.389	0.449	0.954	-1.69	0.92
		Lg. Grid	-0.556	0.449	0.818	-1.86	0.75
		RSVP	-0.5	0.449	0.875	-1.8	0.8